Inversion of Reflection Seismograms by Differential Semblance Analysis: Algorithm Structure and Synthetic Examples

William W. Symes
Michel Kern

July, 1992

TR92-03

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
JUL 1992		**00-00-1992 to 00-00-1992**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Inversion of Reflection Seismograams by Differential Semblance Analysis: Algorithm Structure and Synthetic Examples	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Computational and Applied Mathematics Department ,Rice University,6100 Main Street MS 134,Houston,TX,77005-1892**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**		**75**	

Inversion of reflection seismograms by differential semblance analysis: algorithm structure and synthetic examples

William W. Symes
Michel Kern

The Rice Inversion Project
Department of Computational and Applied Mathematics
Rice Unversity
Houston, Texas 77005

August 4, 1992

Abstract

Seismograms predicted from acoustic or elastic earth models depend very nonlinearly on the long wavelength components of velocity. This sensitive dependence demands the use of special variational principles in waveform-based inversion algorithms. The differential semblance variational principle is well-suited to velocity inversion by gradient methods, since its objective function is smooth and convex over a large range of velocity models. An extension of the adjoint state technique yields an accurate estimate of the differential semblance gradient. Nonlinear conjugate gradient iteration is quite successful in locating the global differential semblance minimum, which is near the ordinary least squares global minimum when coherent data noise is small. Several examples based on the 2D primaries-only acoustic model illustrate features of the method and its performance.

Acknowlegement: This work was partially supported by the National Science Foundation (DMS8905878), the Office of Naval Research (N00014-89-J-1115), the Defense Advanced Research Projects Agency and the Air Force Office of Scientific Research (AFOSR 90-0334), the Texas Geophysical Parallel Computation Project, and The Rice Inversion Project. TRIP Sponsors for 1992 are Amoco Research, Conoco Inc., Cray Research, Earth Modeling Systems, Exxon Production Research Co., and Mobil Research and Development Corp.

Introduction

This paper presents an implementation of a *differential semblance optimization* algorithm for estimation of velocities and reflectivities from multioffset reflection seismograms. Differential semblance combines concepts from nonlinear least-squares inversion, migration velocity analysis, and travel time tomography. It poses the estimation problem as a variational principle of least-squares type. This principle is defined in an expanded space of experiment-dependent or *incoherent* models, which contains the normal model space as a proper subset. A gradient-based local optimization algorithm solves this problem efficiently, whereas other least-error formulations of multioffset waveform inversion appear to require global optimization algorithms such as Monte Carlo search or simulated annealing (Mosegard and Tarantola 1991 [21], Stoffa and Sen 1991 [27], Scales et al., 1991 [26]). Nonetheless, the solution of the differential semblance optimization problem closely approximates the solution of the straightforward mean-square best-fit formulation of velocity estimation when the amount of coherent noise in the data is small. Thus differential semblance optimization may be regarded as an *infeasible point method* for ordinary least-squares inversion.

The next section introduces the differential semblance approach as a variant of nonlinear least-squares inversion, in general terms. The following two section present the detailed formulation of differential semblance optimization for primaries-only 2D acoustics, and describe the gradient calculation and optimization, respectively. In fourth section several simple layered and nonlayered synthetic examples illustrate the behaviour of the method. The algorithm described here applies *a priori* to multidimensional models. The translational symmetry of data from layered models allows the extraction of all kinematic and dynamical information from a small part of the data, however. We take advantage of this redundancy to minimize the computational expense of inversion of layered models, without changing the algorithm. The inversion examples presented below are all layered for this reason. We also exhibit differential semblance gradients for nonlayered models, which confirm the sensitivity of the objective function to lateral velocity heterogeneities.

The paper ends with a discussion of migration velocity analysis, travel time tomography, and least-squares inversion and shows how differential semblance optimization is closely related to each of these techniques. An appendix provides a detailed derivation of the gradient calculation.

The differential semblance concept

Differential semblance optimization is a variant of nonlinear least squares inversion, in which an earth model is sought to minimize the difference between predicted and measured data. Least squares inversion accomodates a variety of physical descriptions of seismic wave propagation and possesses an elegant information-theoretical justification (Tarantola 1987) [41]. As a result the technique has received considerable attention in the last years.

One of the first results of numerical work on least squares was the discovery that the mean square misfit function is severely nonconvex in the long wavelength or smooth components of velocities (Kolb et al. 1986 [17], Gauthier et al. 1986 [14], Symes and Santosa 1989 [24]). As a result most recent work has either used conventional velocity analysis or travel time tomography to adjust the smooth part of the velocity (Helgesen and Kolb 1989 [16], Noble 1992 [22]) or employed global nonconvex optimization technology such as simulated annealing or genetic algorithms (Cao et al. 1990 [5], Mosegard and Tarantola 1991 [21], Stoffa and Sen 1991 [28], [27] Scales et al. 1991 [26]). The latter techniques appear to require a very large number of simulations (10^5 - 10^6 are typical numbers). They depend for their success on the presumption that a "significant" part of model space will have been sampled, but no criteria for adequate sampling appear to be available.

The root cause of the nonconvexity of the least squares objective is the difficulty of fitting all gathers in a multi-gather (i.e. multi-experiment) data set simultaneously, unless the velocity (i.e. the kinematical aspect of the earth model) is very nearly correct. This observation holds for all types of data gathers (common source, common receiver, common offset, plane wave,...). On the other hand fitting a *single* gather is a relatively easy task.

This reasoning suggests the introduction of an experiment-dependent model space. Since the actual earth is not experiment-dependent, a suitable objective function for inversion should be designed to minimize the variance of the model with experiment. Accordingly we propose an objective function containing:

- a least squares (or other convenient) data fitting term;
- a *differential semblance* term penalizing deviations between the models for *neighboring* experiments.

Now such an objective function is just as nonconvex as is is the usual least squares objective, since the latter is contained in the former. However the analysis of this pathology also suggests a way out. Recall that the long wavelengths of the velocity (i.e. p-wave velocity, or p-wave and s-wave velocities if s-wave data is present, etc.) have a markedly different effect on the data than do all other model parameters, such as short-wavelength velocity fluctuations,

impedance contrasts, etc. This dichotomy of data influence accounts is the main source of nonconvexity of the least squares objective, mentioned above. Differential semblance optimization separates the two classes of parameters, viewing the long wavelengths of the velocity as "control" variables and the rest of the model as a "state" variable. The main result of the theoretical and numerical investigation of differential semblance is:

> *If the differential semblance objective is minimized first over the "state" variables, the remaining function of the "control" variables is smooth and convex in a large domain, so can be minimized effectively by gradient-based optimization methods.*

We shall refer to the long wavelengths of the velocity, i.e. the "control" variables, as "the velocity", and the rest of the model, i.e. the "state" variables, as "the reflectivity". Thus the two-stage approach just suggested amounts to inverting for reflectivities given velocities, then stuffing the inverted reflectivities into the differential semblance objective and minimizing the resulting function of velocities. The separation into the two classes can be either implicit or explicit. In this paper we explicitly separate the model into velocity and reflectivity *via* linearization, with the velocity used as the background model. That is, we will adopt the primaries-only approximation. For an example of differential semblance with implicit separation and fully nonlinear modeling (inclusive of multiple reflections), see Symes 1991a [34].

Figure 1 illustrates the reasoning behind our main result. It caricatures inverted reflectivities for correct and incorrect velocities, and displays slices through the inverted reflectivity hypercube at two adjacent shot locations. Inverted reflectivities from neighboring shots are similar in phase content even when the velocity model is wrong (provided that the experiment parameter, i.e. shot location, is sampled finely enough). Therefore their differences change slowly when the velocity model is changed. This accounts for the smooth behaviour of the differential semblance objective.

Symes 1991a [34] and Symes and Carazzone 1991 [37] give an account of an earlier implementation of differential semblance optimization, specifically formulated for plane wave data and layered earth models. This earlier version also extracted reasonable satisfactory estimates of velocity and reflectivity from carefully selected and prepared field data sets.

The implementation presented here was introduced and analysed in Symes 1991 [33], [35] and Symes 1992 [36]. It applies in principle to virtually any class of earth models and presentation of the data. We will limit our attention to 2D acoustic models and shot gather (common source) data sets. As mentioned above we use first-order perturbation theory to view the rest of the model (the reflectivity) as a perturbation about the long wavelengths of velocity. We regard the data as a measurement of the perturbational (scattered) field. Physically,

this approximation amounts to neglect of multiply reflected energy. It appears to be roughly correct for some data sets, admissible after preprocessing for others, and simply inadequate for yet others. We will also neglect density variations, and regard the source as known in both waveform and radiation pattern. These last two assumptions are inessential to the method - both can be dropped at the expense of more complex implementation. In particular, the source wavelet and directivity can be estimated from the data itself as part of the inversion (Lewis 1989 [20]). Since reflection amplitudes in all real data sets are affected by both density fluctuations and source details, we have applied the constant density implementation to synthetic data only.

Differential semblance for 2D primary reflection acoustics

This section defines a mathematically precise version of differential semblance optimization for 2D primaries only constant density acoustics, summarizes some of its properties, and outlines basic computational choices made in our implementation.

We will use a version of the 2D acoustic constant-density primaries-only model of seismic wave propagation. This model assumes that:

- the earth is two-dimensional, and occupies the half-space $\{z > 0\}$;
- the acoustic pressure field vanishes on the surface of the earth (i.e. on $\{z = 0\}$);
- density variations can be neglected;
- multiply reflected energy can be neglected;
- sources are temporally and spatially localized, the source characteristics are known, and differ from shot to shot only by position.

The model couples the fields

- $p(x, z, t; x_s)$ = (scattered) pressure field
- $p_0(x, z, t; x_s)$ = reference pressure field
- $v(x, z)$ = (smooth) velocity field
- $r(x, z)$ = (rough) reflectivity field, i.e. relative velocity perturbation ($\sim 2\delta v(x, z)/v(x, z)$)
- $f(x, z, t; x_s)$ = source (divergence of a body force field)

Here x_s ranges over a set of source locations. The fields are connected through the set of coupled wave equations

$$(\frac{1}{v^2}\frac{\partial^2}{\partial t^2} - \nabla^2)p_0 = f \tag{1}$$

$$(\frac{1}{v^2}\frac{\partial^2}{\partial t^2} - \nabla^2)p = 2r\nabla^2 p_0 \tag{2}$$

Equation 1 is the ordinary constant-density acoustic wave equation for the reference (incident) pressure field p_0. First-order perturbation theory plus a little

algebra produces equation 2 for the perturbation or scattered field p in response to a perturbation δv of v.

So long as v is smooth, i.e. has a characteristic length scale of several wavelengths, the reference field p_0 contains only the direct wave and refracted waves, and is essentially devoid of reflections. This well-known result follows from geometric optics (Friedlander 1958 [13]; for this application see Claerbout 1985 [7], for example). The relative perturbation $r = \delta v/v$ turns up naturally, as a secondary source multiplying the Laplacian of p_0 on the right-hand side of the wave equation for p. Since r is responsible for all reflections in this model, it is natural to call it the *reflectivity*.

To complete the specification of the fields initial and boundary conditions are necessary. We assume that the wave equations hold in a spatial region R and for a time interval $t < t_{max}$, and that the fields are quiescent for negative times:

$$\left.\begin{array}{c} f(x,z,t;x_s) \\ p_0(x,z,t;x_s) \\ p(x,z,t;x_s) \end{array}\right\} \equiv 0, t < 0$$

Ideally R is the half-space $\{z > 0\}$. For computational purposes R must be a finite domain, which we take to be the rectangle $\{x_{min} < x < x_{max}, 0 < z < z_{max}\}$. We assume zero pressure on the surface $\{z = 0\}$:

$$\left.\begin{array}{c} p_0(x,0,t;x_s) \\ p(x,0,t;x_s) \end{array}\right\} \equiv 0$$

It is commonplace to employ so-called absorbing boundary conditions on the other sides of the rectangle to simulate wave propagation in an infinite domain (see eg. Enquist and Majda 1977 [11]). However such conditions implicitly assume that no reflections occur outside R; if this assumption is incorrect, some part of the data may not be explicable by the model. For an example of this sort of effect in an inverse problem (see Noble 1992 [22]). Instead we prefer to bear the extra computational expense of putting the boundaries of R far enough away from the source that no reflection can arrive in the receiver array from the boundary in the recording interval $\{0 < t < t_{max}\}$. That is, we assume

- that p_0 and p are required to vanish on *all* sides of R for all t;
- that the source $f(x,z,t;x_s)$ vanishes outside of a small region near the *source location* x_s, z_s, and that the source depth z_s is the same for all sources;
- that for each source location x_s a set of receiver locations x_r, z_r ("cable") is given, and that the receiver depth z_r is the same for all source locations x_s and receiver offsets $x_r - x_s$;

- that R is sufficiently large that p_0 takes the same values as it would if calculated in the half-space, at all points x_r, z_r of the cable, for all x_s, and for $0 < t < t_{max}$.

In practice it is easy to check that R is sufficiently large, using an average value of the slowness $1/v$.

We solve the wave equations 1, 2 numerically using a finite difference method of fourth order in space and second order in time (Dablain 1986 [8], Levander 1988 [19]). Since the boundary conditions specify the vanishing of the fields p_0 and p on the boundary of R, the method of images gives numerical boundary conditions (i.e. one-sided difference stencils near the boundary) of the same accuracy as the interior scheme.

The *simulated primary reflection data set* consists of the samples of the scattered field p for various source locations x_s, as x ranges over the corresponding receiver coordinates x_r and $0 < t < t_{max}$. Since this data set depends functionally on the fields v and r, we write

$$F[v, r] = \{p(x_r, z_r, t; x_s)\} \tag{3}$$

We call F the *forward map* of this model. Since the receiver positions do not necessarily fall on the finite difference gridpoints, numerical implementation of this map requires an interpolation onto receiver positions.

Note that

- F is *linear* in r, since r appears linearly on the right hand side of 2;
- F is extremely *nonlinear* in v.

The first property is obvious, and the second is by now well established (Claerbout 1985 [7], Tarantola 1986 [40], Santosa and Symes 1989 [24], Symes 1992 [36]). The consequences for waveform inversion are very important. First, given v, it is relatively easy to estimate r. Essentially, estimation of r given v requires a before stack migration with correct output amplitudes. Both direct methods of Kirchhoff type (Beylkin 1985 [1], Bleistein 1987 [3]) and iterative methods based on modeling (Bourgeois et al. 1989 [4], Noble 1992 [22]) work reasonably well. However estimation of v and r jointly is very difficult, because of the second property. For example the mean square misfit function

$$J_{LS}[v, r] = \frac{1}{2}\|F[v, r] - F_{data}\|^2 \equiv \frac{1}{2}\sum_{x_s, x_r, t} |p(x_r, z_r, t; x_s) - p_{data}(x_r, t; x_s)|^2 \tag{4}$$

is highly non-quadratic and does not admit global minimization *via* gradient (local) optimization methods. As mentioned in the introduction, several authors have suggested stochastic minimization of J_{LS}.

For the sake of efficiency, and for other reasons to be discussed later, we prefer to modify the problem and the objective function. Our modification proceeds in two steps.

The first step follows from the behaviour of J_{LS} when only one source (location) is used. Then the minimization of J_{LS} is actually quite straightforward: minimization over r (the quadratic problem) yields a value essentially independent of v. Thus minimization of J_{LS} for a single shot is *easy*, and does not constrain v at all. The difficult nature of minimizing J_{LS} in general stems from the necessity of fitting data from many shot locations simultaneously, with a single model $[v, r]$. This circumstance suggests *relaxing* the minimization problem by allowing the model to depend on the source location. Since v is unconstrained by independent fits to the various single-source data sets, only r actually need be allowed to vary with x_s:

$$v = v(x, z)$$
$$r = r(x, z; x_s)$$

Now it will be easy to achieve essentially zero residual value of the functional defined in 4, but at the cost of losing control completely over v. Also the estimates of r will actually depend on x_s, which is nonsense since r is supposed to be part of an earth model. The second step recovers control over v, in a calibrated way, by *penalizing* the dependence of r on x_s. For reasons discussed in a general way in the introduction, and justified mathematically in Symes 1991c [35] and 1992 [36], we choose to add *differential* penalty term:

$$J_{DS}[v, r] = \frac{1}{2}\{\|F[v, r] - f_{data}\|^2 + \sigma^2 \|\partial r/\partial x_s\|^2\} \tag{5}$$

This is the *differential semblance objective function.* Making the first term small forces some level of fit-to-data, independently for each source location. Making the second term small forces the x_s-dependent reflectivities to resemble each other. The two goals can be achieved simultaneously only when the velocity is kinematically correct. The weight parameter σ controls the amount of penalty applied for incoherence of r, and thus interpolates between totally independent data fitting (4 for $r = r(x, z; x_s)$, equivalent to 5 for $\sigma = 0$) and the ordinary least-squares principle (4 for $r = r(x, z)$, equivalent to 5 for $\sigma \to \infty$).

Now J_{DS} is no more smooth or convex than is J_{LS}, since it essentially contains J_{LS} as its first term. However, as argued in the introduction J_{DS} can still be minimized effectively *via* local optimization algorithms, provided that the optimization is divided into two stages:

1. first carry out the (quadratic) minimization over r, as a result of which r becomes a function of v: $r = r[v]$ ("the inverted reflectivity");
2. then minimize the residual $J_{DS}[v, r[v]]$ of step 1 over v.

Two elements of this scheme cooperate to make it effective. First, only inverted reflectivities from *neighboring* source locations are compared. These are similar even when the velocity field v is substantially wrong. Numerical implementation replaces the partial derivative by a divided difference, of course:

$$\frac{\partial r}{\partial x_s} \rightarrow \frac{r(\cdot,\cdot,x_s+\Delta x_s) - r(\cdot,\cdot,x_s)}{\Delta x_s}$$

So long as the sample rate Δx_s is small enough relative to the dominant spatial wavelength and the maximum traveltime error, this difference will not involve any cycle-skipping.

Second, *inverted* reflectivities are compared. These really should all equal the target (x_s-independent) reflectivity if the velocity is chosen correctly. In particular the $r[v](\cdot,\cdot,x_s)$ should largely be free of amplitude artifacts, apart from those unavoidable errors caused by limited data aperture. Therefore a simple divided difference, as above, should be sufficiently robust to act as a velocity indicator.

Apart from these intuitive arguments, we have given some mathematical analysis in Symes 1991c [35] and 1992 [36] which shows that $J_{DS}[v, r[v]]$ is smooth and convex in a large region of velocity models v, for proper choice of the parameter σ. Also, the numerical evidence is strongly in favor of this conclusion, as will be reviewed below.

Now suppose that the data F_{data} is *consistent*: for some model $[v^*, r^*]$,

$$F_{data} = F[v^*, r^*] \tag{6}$$

Here r^* is supposed to be x_s-independent. Then

$$J_{LS}[v^*, r^*] = 0 = J_{DS}[v^*, r^*]$$

so that both objective functions share the same global minimum. In other words, global minimization of J_{LS} is accomplished by global minimization of J_{DS} in the consistent (noise-free) data case. Thus one can view differential semblance optimization as an *infeasible point* method for least squares inversion.

We also have some mathematical and considerable numerical evidence that the minimization of J_{DS} is *stable* in a suitable sense. That is, suppose that 6 fails but the two sides are close in the mean square sense (i.e. the RMS data noise is small). Then the minimizer v of $J_{DS}[v, r[v]]$ is *kinematically close* to v^*, in the sense that these generate RMS-close travel time fields from all source and receiver positions. This conclusion appears to continue to hold even for quite large amounts of *incoherent* data noise.

Algorithmic details

Evidently the two-stage procedure for minimizing J_{DS} outlined in the last section really requires an algorithm for minimizing the function of velocity alone:

$$\tilde{J}[v] \equiv J_{DS}[v, r[v]]$$

where

$$r[v] = argmin_r J_{DS}[v, r].$$

The differential semblance objective function J_{DS} is defined by

$$J_{DS}[v, r] = \frac{1}{2}\{\|F[v, r] - F_{data}\|^2 + \sigma^2\|\partial r/\partial x_s\|^2 + \lambda^2\|Wr\|^2\}$$

Here we have added a *regularizing term* with weight λ and operator W, to ensure that the inner minimization over r is well-behaved. Naively one could take $W = I$, and this works to some extent. A better choice is $W = \nabla_{x,z,x_s}$, as discussed below.

We shall use the nonlinear conjugate gradient algorithm as presented in Fletcher 1980 [12] to minimize $\tilde{J}$. This algorithm produces a sequence $v_i, i = 0, 1, 2, ...$ according to the rules

1. Initialization:

$$v_1 = v_{initial}, s_0 = 0, \beta_0 = 0$$

2. for $k = 1$ until convergence do:

$$s_k = -grad\tilde{J}[v_k] + \beta_{k-1}s_{k-1}$$

$$v_{k+1} = v_k + \alpha_k s_k$$

The step update parameter β_k is an algebraic combination of inner products of the gradients $grad\tilde{J}[v_{k-1}]$ and $grad\tilde{J}[v_k]$ (Polak-Ribiere formula). The step α_k is chosen by a *line search*, which is a safeguarded version of the secant minimization method for functions of a single variable. It involves algebraic combinations and tests using the function value $\tilde{J}[v_k + \alpha s_k]$ and the slope $< s_k, grad\tilde{J}[v_k + \alpha s_k] >_v$. Here $< \cdot, \cdot >_v$ is the symbol we shall use for the inner product in the Hilbert space of velocity changes. The same inner product is used in the Polak-Ribiere formula. Convergence is tested by decrease of the gradient norm:

$$\text{convergence} \Leftrightarrow \|grad\tilde{J}[v_k]\|_v \leq \varepsilon\|grad\tilde{J}[v_{initial}]\|_v$$

Here $\|\cdot\|_v$ is the norm associated with the inner product $< \cdot, \cdot >_v$.

To carry out this program, we must settle on methods for computing

1. the inner product $< \cdot, \cdot >_v$;
2. the value $\tilde{J}[v]$;
3. the gradient $grad\tilde{J}[v]$.

The subsections to follow describe each of these calculations.

The Hilbert space structure of velocity updates

The theory underlying all of the velocity analysis methods mentioned in the introduction, including differential semblance, depends on the *separation of scales* between velocity in the one hand and reflectivity on the other. In many cases the large-scale nature of velocity has been enforced through parameterization, by selecting velocities from a fixed low-dimensional space of smooth functions. Obviously such a constraint is rather arbitrary, and it is far from obvious that a choice of parameterization could be made which yields large-scale structure and also reproduces travel times with the necessary fidelity in every situation. If the initial velocity estimate $v_{initial}$ is smooth (for instance constant or constant gradient), then control of the length scales in the space of velocity updates is the important issue. We prefer to allow the velocity updates δv the same formal degrees of freedom as the other grid functions, and enforce *only* smoothness through the choice of norm or inner product. In this approach the Hilbert space structure of velocity updates becomes extremely important.

We use a discrete analogue of the *Sobolev scale of Hilbert norms*. These are based on the anisotropic discrete Laplace operator

$$\delta v_{i,j} \mapsto w_x^2 \frac{\delta v_{i+1,j} + \delta v_{i-1,j} - 2\delta v_{i,j}}{\Delta x^2} + w_z^2 \frac{\delta v_{i,j+1} + \delta v_{i,j-1} - 2\delta v_{i,j}}{\Delta z^2} \equiv (L\delta v)_{i,j}$$

With suitable boundary conditions this operator is diagonalized by the Fourier transform, so it is possible to compute arbitrary powers of L very efficiently. The sth Sobolev norm is given by

$$||\delta v||_s \equiv ||\Lambda^s \delta v||$$

where the norm without the subscripts is the ordinary L^2 norm

$$||u||^2 = \Delta x \Delta z \sum_{i,j} |u_{i,j}|^2$$

and

$$\Lambda \equiv (I - L)^{\frac{s}{2}}$$

The operator Λ multiplies the (k, l) discrete Fourier coefficient of δv by $(1 + w_x^2 k^2 \Delta x^2 + w_z^2 l^2 \Delta z^2)^{\frac{1}{2}}$ for small wavenumbers $k\Delta x, l\Delta z$. Thus the power s

determines the rate at which high frequencies are emphasized, and the weights w_x, w_z determine the corner frequencies in each direction.

For reasons to be explained below, we constrain the velocity updates to vanish in a surface layer containing the source:

$$\delta v(x,z) \equiv 0 \text{ for } 0 < z < z_{min}$$

The boundary conditions imposed on L in the definition of Λ should include $u = 0$ at the grid level just below z_{min}. The other three boundary conditions are less important.

Besides defining the inner products in the conjugate gradient formulae, the operator Λ^s plays a critical role in the definition of the adjoint operators, and thereby on the gradient. Indeed, suppose that A is some other operator taking velocity updates δv to some other sort of grid functions. We use the Sobolev s norm on the space of velocity updates, and the ordinary (L^2) norm on the other space of grid functions. We assume that the ordinary (L^2) transpose or conjugate operator A^T is available. That is, A^T is the operator which would be computed by transposing the matrix of A, if one could access the matrix elements of A (out of the question in almost all examples of interest to us!). It is defined by the condition

$$< A\delta v, \phi >=< \delta v, A^T\phi >$$

for all suitable grid functions $\delta v, \phi$. On the other hand the adjoint for the Sobolev s norm on δv is defined by the similar formula

$$< A\delta v, \phi >=< \delta v, A^*\phi >_s=< \Lambda^s\delta v, \Lambda^s A^*\phi >=< \delta v, \Lambda^{2s}A^*\phi >$$

Inspecting these two formulas we see that

$$A^T = \Lambda^{2s}A^*$$

Since Λ^{-2s} is inverse to Λ^{2s}, this is equivalent to

$$A^* = \Lambda^{-2s}A^T \tag{7}$$

From the remark above about the Fourier coefficients we see that if $s > 0$, then Λ^{-2s} suppresses high frequencies, i.e. smooths its argument. Thus

> *the ordinary transpose of an operator is converted to its adjoint in the sense of the Sobolev s-norm by application of a smoothing operator.*

This observation will play a crucial role in explaining the construction of the velocity gradient below.

Evaluation of $\tilde{J}[v]$

Evaluation of $\tilde{J}$ requires solution of the quadratic optimization problem

$$minimize_r J_{DS}[v, r]$$

to produce $r[v]$ as well as the value $\tilde{J}[v] = J_{DS}[v, r[v]]$.

Recall that

$$J_{DS}[v, r] = \frac{1}{2}\{||F[v, r] - F_{data}||^2 + \sigma^2||\partial r/\partial x_s||^2 + \lambda^2||Wr||^2\}$$

The variation or directional derivative of J_{DS} in the direction δr is

$$\begin{aligned} \delta J_{DS}[v, r] &= < D_r F[v, r]\delta r, F[v, r] - F_{data} > \\ &+ \sigma^2 < \partial\delta r/\partial x_s, \partial r/\partial x_s > \\ &+ \lambda^2 < W\delta r, Wr > \end{aligned}$$

Here $< \cdot, \cdot >$ is the L^2 norm over x, z, and x_s (discretely the Euclidean dot product weighted by $\Delta x \Delta z \Delta x_s$). Because of the convention mentioned in the last section, that the boundaries of the computational domain R are either pressure free or are so far away as not to matter, we can assume that r and δr vanish near the boundary of R without loss of generality. Integration (discretely, summation) and definition of the adjoints gives

$$\delta J_{DS}[v, r] =< \delta r, D_r F[v, r]^*(F[v, r] - F_{data}) - \sigma^2\partial^2 r/\partial x_s^2 + \lambda^2 W^T W r >$$

from which we read off immediately

$$grad_r J_{DS}[v, r] = D_r F[v, r]^*(F[v, r] - F_{data}) - \sigma^2\partial^2 r/\partial x_s^2 + \lambda^2 W^T W r$$

Setting the gradient to zero gives the *normal equation*. Because F is linear in r,

$$D_r F[v, r]\delta r = F[v, \delta r]$$

whence the *adjoint operator*

$$M[v] \equiv D_r F[v, r]^*$$

is independent of r and the normal equation is linear in r. With proper choice of regularization, this positive definite symmetric linear system is necessary and sufficient for optimality of r. It can be re-written as

$$N[v]r = M[v]F_{data} \tag{8}$$

where

$$N[v]r \equiv M[v]F[v, r] - \sigma^2\partial^2 r/\partial x_s^2 + \lambda^2 W^T W r$$

is the *normal operator.*

Simply to evaluate the both sides of the normal equations requires algorithms for the action of $N[v]$ and $M[v]$. Our approach to solving the normal equations will also make repeated use of such algorithms. The normal operator $N[v]$ is composed of $M[v]$, $F[v,r]$, $\partial/\partial x_s$, and the regularizing operator W, which is either a scalar operator or made up out of derivatives in our current implementation. The derivatives are approximated by finite difference operators in obvious ways, and we described the calculation of $F[v,r]$ in the last section. Thus it only remains to specify an algorithm for the action of the adjoint operator $M[v]$ on a field of "residuals" $\phi(x_r,t;x_s)$. This algorithm is the so-called adjoint state method, and has been explained a number of times in the literature; see especially Lailly 1983 [18], Tarantola 1984 [39]. In brief,

1. backpropagate the residuals, i.e. solve the final value problem for each x_s:

$$(\frac{1}{v^2}\frac{\partial^2 w}{\partial t^2} - \nabla^2 w)(x,z,t;x_s) = \sum_{x_r} \phi(x_r,t;x_s)\delta(x-x_r,z-z_r) \qquad (9)$$

$$w \equiv 0, t > t_{max}$$

2. crosscorrelate the backpropagated field w with the incident field p_0:

$$M[v]\phi = -\frac{2}{v}\int dt w \nabla^2 p_0$$

The field w satisfies the same (homogeneous) boundary conditions as do p_0 and p. Note that the final values of p_0 can be saved, and the field reconstructed backwards in time, since the wave equation is time-reversible. Thus the output of the crosscorrelation can be accumulated as the calculation proceeds, with tremendous savings in storage. Note that the output $M[v]\phi$ in this context is x_s-dependent. In fact, $M[v]$ is exactly a prestack migration operator. The analogous operator for ordinary least-squares inversion is computed in exactly the same way, but includes a final summation (stack) over x_s.

Because of the size of the normal equation 8, it is most conveniently solved by a *polynomial iteration.* While the conjugate gradient family is favored in most similar work, and is optimal in a sense we will not make precise, we prefer to use *Chebyshev iteration* for reasons that will become clear later.

Chebyshev iteration involves four distinct steps:

1. choosing an *inversion level*;
2. estimating the spectral bound and scaling the normal operator;
3. computing the Chebyshev coefficients and estimating the necessary number of iterations;

4. application of the Chebyshev polynomial by recursive application of the normal operator.

A good reference on classical Chebyshev iteration as applied to discretized elliptic boundary value problems is Varga 1962 [43]. Stork and Clayton 1991 and 1992 [31], [32] reported the use of Chebyshev accelerated Richardson iteration applied to reflection tomography. They pointed out the the advantages of the control provided by Chebyshev iteration over the portion of the spectrum inverted. Their algorithm is somewhat different than ours. In particular they are able to normalize their operator so as to be able to apply Chebyshev acceleration directly, whereas we must compute an approximate spectral bound. We were unable to locate a description of the modifications necessary to handle the present problem, so we will give a fairly detailed description of the algorithm.

Choosing the inversion level

Chebyshev iteration reduces the error and residual components of a symmetric linear system to a specified level relative to their original values over a specified part of the operator spectrum. For positive definite problems like the normal equations it is natural to specify a lowest eigenvalue at which the reduction will be achieved, and to give its value relative to the largest eigenvalue (spectral bound) λ_{max} of $N[v]$. Therefore it is necessary to choose the *error reduction factor* ε and the *inversion level* γ. The iteration will be designed so that if $\gamma\lambda_{max} \leq \lambda \leq \lambda_{max}$ and λ is an eigenvalue of $N[v]$ then the corresponding component of the *normal residual* $M[v]F_{data} - N[v]r$ will be reduced by the factor ε (at least). Evidently a good choice of inversion level depends on the spectral structure of the data, as does the choice of error reduction factor. Stork and Clayton 1991, 1992 [31], [32] suggest that in tomographic inversion γ should be chosen quite small, a typical figure being 0.04.

Estimating the spectral bound and scaling the normal operator

Of course the spectral bound is itself unknown *a priori*. It represents the natural scale of the problem, so is valuable information in its own right. Fortunately underestimates are a by-product of the iteration, in the form of *Rayleigh quotients* of the reflectivity updates δr_k, at the cost of a couple of extra inner products:

$$RQ_k \equiv \frac{< \delta r_k, N[v]\delta r_k >}{< \delta r_k, \delta r_k >}$$

Any number strictly larger than the maximum of these estimates can be used in the place of the actual spectral bound λ_{max}. The additional steps needed to make use of these adaptive spectral estimates are as follows:

1. choose a "fudge factor" $\alpha > 1.0$;

2. replace the error reduction factor ε by

$$\varepsilon_{est} = \frac{\sqrt{\alpha - 1}\varepsilon}{1 + \sqrt{\alpha - 1}}$$

3. initialize the spectral bound λ_{est} by

$$\lambda_{est} = \alpha RQ_0$$

   or by

$$\lambda_{est} = max(\lambda_{est}, \alpha RQ_0)$$

   if a previous estimate is available.

4. if at step k, $RQ_k > \lambda_{est}$, replace λ_{est} by αRQ_k, recompute ε_{est} and all quantities dependent on it, and restart the iteration.

It is simple to show that the adaptive iteration, thus safeguarded, produces final error reduction of at least ε.

In practice we have found that the spectral bound estimated in this way is quite stable. Initially, one or two (or rarely three) restarts are necessary. Subsequent similar inner inversions with updated velocities, as occur in the outer inversion loop, can usually proceed with a previous λ_{est}.

Since Chebyshev iteration in its original form applies to operators with spectra lying between -1 and 1, it is in principle necessary to scale and translate $N[v]$, i.e. to replace it by $I - sN[v]$ for a suitable scale factor s, to satisfy this condition. We have re-written the algorithm so that this step is carried out implicitly; nontheless the scale factor s is needed. An optimal choice is

$$s = \frac{2}{(1 + \gamma)\lambda_{max}}$$

where in practice λ_{max} is replaced by λ_{est}.

Computing the Chebyshev coefficients and estimating the necessary number of iterations

The necessary coefficients are generated by the following recursions, in which the quantity

$$\beta \equiv \frac{1 - \gamma}{1 + \gamma}$$

figures prominently (see eg. Varga 1962 [43]):

$$\begin{aligned} c_0 &= 1 \\ c_1 &= \frac{1}{\beta} \\ c_{k+1} &= \frac{2}{\beta^2} c_k - c_{k-1}, k = 1, 2, ... \\ \omega_0 &= 0 \\ \omega_1 &= 1 \\ \omega_{k+1} &= 1 + \frac{c_{k-1}}{c_{k+1}}, k = 1, 2, ... \end{aligned}$$

The error reduction factor after k steps is

$$\varepsilon_k = \frac{2q^k}{1 + q^{2k}}$$

where

$$q \equiv \frac{\beta}{1 - \sqrt{1 - \beta^2}}.$$

The simplest way to ensure the required decrease is to compute these numbers until

$$\varepsilon_k < \varepsilon_{est}$$

which give the necessary number of iterations k_{max}.

Application of the Chebyshev polynomial by recursive application of the normal operator

The usual arrangement of Chebyshev iteration is as a three-term recursion - see Varga 1962 [43] for instance. We prefer to arrange the algorithm so that it resembles the usual form of conjugate gradient iteration, as follows:

1. initialize:

$$\begin{aligned} r_0 &= 0 \\ \delta r_0 &= 0 \\ e_0 &= M[v] F_{data} \\ nr_0 &= N[v] e_0 \\ RQ_0 &= \frac{< e_0, nr_0 >}{< e_0, e_0 >} \end{aligned}$$

2. for $k = 0 : k_{max} - 1$

$$\begin{aligned} \delta r_{k+1} &= (\omega_{k+1} - 1)\delta r_k + s\omega_{k+1} e_k \\ r_{k+1} &= r_k + \delta r_{k+1} \\ nr_{k+1} &= N[v]\delta r_{k+1} \\ e_{k+1} &= e_k - nr_{k+1} \\ RQ_{k+1} &= \frac{< \delta r_{k+1}, nr_{k+1} >}{< \delta r_{k+1}, \delta r_{k+1} >} \end{aligned}$$

The final outputs of this process are $r[v]_{est} \equiv r_{k_{max}}$, the estimated reflectivity, and the estimated normal residual

$$\begin{aligned} e[v]_{est} &\equiv e_{k_{max}} \\ &= M[v]F_{data} - N[v]r[v]_{est}, \end{aligned}$$

We then compute the *seismogram misfit*

$$m[v]_{est} \equiv F[v, r[v]_{est}] - F_{data}$$

and finally the estimated value of $\tilde{J}$:

$$\begin{aligned} \tilde{J}[v]_{est} &\equiv J_{DS}[v, r[v]_{est}] \\ &= \frac{1}{2}\{\|m[v]_{est}\|^2 + \|\partial r[v]_{est}/\partial x_s\|^2 + \|Wr[v]_{est}\|^2\} \end{aligned}$$

Computation of the velocity gradient $grad\tilde{J}[v]$

For a perturbation δv of the velocity we obtain the directional derivative

$$D\tilde{J}[v]\delta v = (D_v J_{DS})[v, r[v]]\delta v + (D_r J_{DS})[v, r[v]]D_v r[v]\delta v \qquad (10)$$

We will first assume that we have actually solved the normal equation 8 derived in the preceding pages, and produce in this way a *naive gradient approximation.* Since equation 8 states exactly that the r-derivative of J_{DS} vanishes at $[v, r[v]]$, the second term in equation 10 vanishes. Since the differential semblance and regularization terms are independent of v, we obtain

$$D\tilde{J}[v]\delta v = (D_v J_{DS})[v, r[v]]\delta v =< D_v F[v, r[v]]\delta v, F[v, r[v]] - F_{data} >$$

Evidently we must construct a bilinear operator B so that for arbitrary data-like ϕ, reflectivity-like q, and velocity-like v and δv,

$$< D_v F[v, q]\delta v, \phi >=< \delta v, B[q, \phi] > \qquad (11)$$

Then

$$\begin{aligned} D\tilde{J}[v]\delta v &= <\delta v, B[r[v], F[v, r[v]] - F_{data}]> \\ &= <\delta v, \Lambda^{-2s} B[r[v], F[v, r[v]] - F_{data}]>_s \end{aligned}$$

recalling the discussion of Sobolev adjoints earlier in the section. Thus the gradient of $\tilde{J}$ in the sense of the Sobolev s-norm is

$$grad\tilde{J}[v] = \Lambda^{-2s} B[r[v], F[v, r[v]] - F_{data}] \tag{12}$$

As noted before, Λ^{-2s} is a smoothing operator for positive s.

Of course in practice we must replace the reflectivity $r[v]$ by its estimate $r[v]_{est}$, obtained as explained in the last section, and the misfit $F[v, r[v]] - F_{data}$ by its estimate $m[v]_{est}$. These replacements in 12 yield the *naive gradient estimate*:

$$grad\tilde{J}[v]_{est} = \Lambda^{-2s} B[r[v]_{est}, m[v]_{est}] \tag{13}$$

To complete the description of the naive gradient estimate, it only remains to describe the calculation of B. This calculation is an extension of the adjoint state technique, and is derived in Appendix A. It is very similar to the computation of gradients for functions of migrated images proposed by Chavent and Jacewitz 1990 [6]. The steps are:

1. compute the backpropagated residual field w, i.e. solve the problem 9 as in the calculation of $M[v]$;

2. compute the secondary backpropagated field w_0 satisfying

$$(\frac{1}{v^2}\frac{\partial^2 w_0}{\partial t^2} - \nabla^2 w_0)(x, z, t; x_s) = 2\nabla^2(qw)$$

$$w_0 \equiv 0, t > t_{max}$$

   with homogeneous boundary conditions; w_0 can be stepped backwards in time along with w and the reconstructions of p_0, p in the finite difference scheme;

3. compute the double crosscorrelation

$$B[q, \phi] = \frac{2}{v}\int dt(w(\nabla^2 p + 2q\nabla^2 p_0) + w_0\nabla^2 p_0)$$

Thus the calculation of the naive gradient approximation is a reasonably straightforward extension of the usual gradient calculation from least-squares inversion. Unfortunately the naive gradient 13 is not very accurate as an approximation to the exact gradient 12. We will illustrate this in the next section. The inaccuracy can be large enough to render 13 *useless for optimization of* $\tilde{J}$*!*

In fact, we have neglected the second term in the equation 10. This can be rather large, even when the normal equation is solved rather precisely, because it involves the *derivative of the reflectivity estimate with respect to velocity.* It is the anomalously large size of the derivative of the synthetic data with respect to velocity that led to nonconvexity of the mean square error function. For exactly the same reason - phase perturbation of rapidly oscillating signals - neglect of the second term in 10 leads to large errors.

Our resolution of this problem involves computing the second term in 10 - not all of it, which would be prohibitively expensive, but its *principal part*, which turns out to be available at reasonable cost. The calculation is tied to our method of producing $r[v]_{est}$, namely Chebyshev iteration, and is derived in Appendix B:

1. Carry out the Chebyshev iteration as above to obtain estimates of reflectivity $r[v]_{est}$, normal residual $e[v]_{est}$, and seismogram misfit $m[v]_{est}$;
2. *Rerun* Chebyshev iteration with the normal residual as data, to produce the auxiliary field y:

$$\begin{aligned} y_0 &= 0 \\ \delta y_0 &= 0 \\ f_0 &= e[v]_{est} \\ ny_0 &= N[v]f_0 \end{aligned}$$

For $k = 0 : k_{max} - 1$

$$\begin{aligned} \delta y_{k+1} &= (\omega_{k+1} - 1)\delta y_k + s\omega_{k+1} f_k \\ y_{k+1} &= y_k + \delta y_{k+1} \\ ny_{k+1} &= N[v]\delta y_{k+1} \\ f_{k+1} &= f_k - ny_{k+1} \end{aligned}$$

The set $y[v]_est \equiv y_{k_{max}}$.

3. Compute the synthetic of the auxiliary field:

$$g[v]_{est} \equiv F[v, y[v]_{est}]$$

4. Replace the naive gradient formula 13 by

$$grad\tilde{J}[v]_{est} = \Lambda^{-2s}(B[r[v]_{est} - y[v]_{est}, m[v]_{est}] - B[r[v]_{est}, g[v]_{est}]) \quad (14)$$

Because of the second Chebyshev iteration and the extra calculations of the forward map F and the bilinear operator B, the *corrected gradient* 14 is roughly twice as expensive as the naive gradient 13. Despite the added expense, we advocate its use because

- we are able to give a mathematical proof that the corrected gradient error, i.e. the difference between 12 and 14, tends to zero as $k_{max} \rightarrow \infty$;
- in many numerical examples, some of which are reviewed below, the corrected gradient has proven *far* more accurate than the naive gradient, for modest k_{max}. In fact usually the corrected gradient with k_{max} iterations is a great deal more accurate than the naive gradient with $2k_{max}$ iterations, so the correction also appears to be *efficient*.

We are not so far able to prove gradient convergence for any correction scheme if Chebyshev iteration is replaced by conjugate gradients. Also a limited amount of numerical evidence suggests that conjugate gradient derived gradients are really not corrected by these tricks. Apparently the constant polynomial coefficients of Chebyshev iteration actually make the estimated $\tilde{J}$ smoother. In any case this partly explains our preference for Chebyshev iteration, for this application.

We have now explained how all pieces of the nonlinear conjugate gradient iteration are calculated. We proceed to examine some numerical results:

Numerical investigation

We present several numerical examples of differential semblance optimization applied to synthetic data sets. The examples fall into two groups. The first group concerns estimation of a layered model. We make use of the observation that the translational symmetry of layered data gives full information on moveout from only two neighboring shot records. The use of a two shot "line" minimizes the expense of each trial, and allowed us to perform much more experimentation than would have been possible with larger simulations. The second group of examples tests the ability of differential semblance to detect strong lateral heterogeneity by exhibiting gradients of the objective $\tilde{J}$.

All examples presented here share the same source and receiver parameters. The "cable" consisted of 34 point receivers spaced 50 m apart, 12 m beneath the free surface. Near offset is 150 m, far offset 1800 m. The source is also punctual and isotropic, and is located at 8 m depth. Shot spacing was 96 m. The source waveform was a Ricker wavelet with 15 Hz peak frequency. The sample rate was 4 ms, and 2 s of data were used. A mute was applied to remove direct and postcritical energy.

The finite difference grid interval was 16 m in both directions. For the velocity ranges used, this choice gave roughly 5 gridpoints per wavelength at minimum velocity and peak frequency. Some evidence of numerical dispersion from the (4,2) finite difference scheme is visible in the synthetics. However numerical dispersion did not appear to be so severe as to change the character of the results.

Layered model

The target layered model appears in Figure 2, and the two-shot data set in Figure 3. Note that the water bottom event is mostly removed from the data by the (very crude linear) mute, and so the zone from the surface to roughly 650 m is essentially quiet.

An effect entirely peculiar to 2 shot data sets and layered media is that the actual value of the differential semblance weight σ becomes essentially irrelevant. Figure 4 shows the curves generated by sampling the objective function $\tilde{J}$ along the line from the constant velocity profile ($v \equiv 1.5$ m/ms) and the target profile, for several values of σ. All curves are essentially quadratic. In the experiments we used $\sigma = 10.0$. We also used a very small amount of damping (0.1 % of the maximum eigenvalue estimate) to enhance convergence of the Chebyshev iteration for reflectivity.

We emphasize that the eventual independence from σ of the objective function is characteristic only of two-shot data sets and layered media. In effect,

for large σ the two reflectivity gathers are forced to be the same, and the DSO objective is essentially identical to the output least squares objective after minimization over reflectivity. Since the two source positions are close, the inverted reflectivity from one data gather would yield a slowly varying error in predicting the other data gather, even when the velocity is dramatically wrong. Therefore the same is true of the reflectivity optimized for both data gathers simultaneously. Otherwise put, since the source positions are close, no "cycle skipping" occurs: for this configuration, output least squares is effective, provided that optimization proceeds in the two-stage fashion we have devised for DSO! As soon as more distance between shot points is involved, as is the case when the number of shots becomes significantly larger, the differential semblance objective depends quite strongly on σ, exhibiting convexity for small σ and approaching the (quite nonconvex) output least squares objective for large σ.

For the inner products in the conjugate gradient formulas, and to smooth the gradient estimate as noted before, we used the Sobolev 2-norm on the space of velocity updates. We set the anisotropic weight parameters in the Laplacian so that the velocity gradient calculation would suppress Fourier components with wavelengths of less than 400 m to 5% or less of their size in the unsmoothed gradient.

For this set of experiments we defined the Chebyshev polynomial (used in the inner optimization for reflectivity) to suppress error components in the normal residual above 2% of the (estimated) maximum eigenvalue to at most 2% of their initial values. This choice implied 19 inner iterations. Parameters for the (outer) nonlinear conjugate gradient iteration over velocity were as advised in Fletcher 1980 [12]. In all cases we used the corrected gradient construction explained in the last section.

The result of four nonlinear conjugate gradient steps is displayed in Figure 5. While the general trend of the velocity to about 1200 m is correct, the details are not reproduced at all. Nonetheless the estimated velocity is kinematically correct. This is evidenced by the comparison of the inverted reflectivity (for shot 1 - shot 2 is similar) shown in Figure 6a with the inverted reflectivity at the correct velocity (with the same iteration parameters) in Figure 6b. The two are essentially equally "flat", whereas the inverted reflectivity at constant velocity (Figure 6c) exhibits considerable residual moveout. Because of the translational invariance of data from layered models, an individual reflectivity gather is equivalent to the (inversion analogue of) a coherency or common image panel, as used in migration velocity analysis (eg. Versteeg and Grau 1991 [45]). Thus its flatness is diagnostic of kinematical correctness of the velocity model.

The RMS data misfit at the end of this exercise was roughly 22%. The predicted data is displayed in Figure 7, the misfit in Figure 8. Evidently all major features are reproduced. While some error energy remains in the major reflector at 1.8 s, much of the remainder appears to be finite difference noise.

To see the improvement possible through further iteration, we performed four more nonlinear conjugate gradient steps. The results are displayed in Figures 9 and 10. Evidently the velocity structure is somewhat improved, though the kinematic effects of this improvement are miniscule.

Table 1 shows the performance of the nonlinear conjugate gradient iteration.

To explore a little bit the effect of one of the parameter choices, we show in Figure 11 the velocity estimates produced by choosing length scales of 300 m, 400 m, and 500 m respectively in the velocity gradient calculation. These estimates resulted from four nonlinear conjugate gradient steps with other parameters set as before. The 300 m result shows considerable detail, all of which is wrong. The reflectivities (Figure 12) are essentially equally flat. The 500 m result shows slightly better depthing of reflectors. These experiments suggest strongly that *only* the kinematic aspects of velocity inversions are to be trusted, never the short-scale details.

Nonlayered model

The second group of experiments used two 35 shot data sets. The goal of this group is to illustrate the ability of differential semblance to detect velocity changes. Accordingly we merely generate gradients rather than attempt full inversions.

The first data set was generated from the layered model used in the first group, so all shots look exactly like those displayed in Figure 3. Figure 13 shows a grey scale plot of the gradient at constant velocity $v \equiv 1.5$ m/ms. Except for edge effects at the ends of the line, this gradient is very nearly layered. Several vertical cross sections appear in Figure 14, illustrating the layered nature of the gradient.

We selected this configuration to illustrate the importance of the gradient correction. We computed the slope of $\tilde{J}$ along the line connecting $v \equiv 1.5$m/ms and the target velocity profile in two ways:

1. by the (Sobolev s-) inner product of gradient with the difference of the velocity profiles;

2. by very careful extrapolated finite difference estimates.

Table 2 compares the results for two different levels of normal residual reduction. Evidently the degree of approximation of the finite difference estimate by the naive gradient (formula 13) barely improves with the threefold reduction in the normal residual, whereas the correct gradient (formula 14) converges superlinearly. We have observed this behaviour in many examples.

To test the ability of differential semblance to detect strong lateral velocity

changes, we added to the model used in the preceding examples a high-velocity near-surface anomaly, roughly circular in shape and with a diameter of about 300 m. A grey-scale plot appears in Figure 15. We used the same (layered) reflectivity model as in the previous examples. The synthetic data (Figure 16) clearly reveals the presence of the anomaly through strongly nonhyperbolic and shot-dependent moveout.

We first computed the gradient at constant $v \equiv 1.5$ m/ms. A grey-scale plot appears as Figure 17. The result is similar to the gradient produced from the layered data (Figure 13). Though evidence of lateral heterogeneity is visible, the actual character of the anomaly is not. A conjugate gradient step would move the velocity estimate in roughly the direction of the layered model.

We then computed the gradient at the layered model used in the preceding experiments. A grey-scale plot appears as Figure 18, and a horizontal cross-section at 300 m depth as Figure 19. The anomaly is now clearly defined. A single conjugate gradient step would add a negative multiple of the gradient to the layered model, producing a high-velocity anomaly in exactly the right place, of approximately correct extent.

Comment on parameters

Evidently differential semblance optimization as formulated in this paper depends on a large number of parameter choices: inversion level and error tolerance for the Chebyshev loop, length scales for the gradient smoothing step, regularization and differential semblance weights, nonlinear conjugate gradient parameters,... Moreover inclusion of flexible constraints reflecting "geological" knowledge, as in Stork and Clayton 1991 [31] for example, while essential for the production of meaningful results, can introduce yet more tuning and penalty parameters. At least partial automation of parameter selection is obviously essential to the eventual success of differential semblance optimization, or indeed any geophysical inversion tool.

Discussion: Migration velocity analysis, travel time tomography, least squares inversion, and differential semblance

Differential semblance optimization is closely related in conception and behaviour to several other approaches to velocity estimation.

Waveform-based velocity estimation methods in widespread use rely on systematic search through a small family of models. A fundamental example of this type is *semblance analysis* of normal moveout corrected common midpoint gathers, as in Taner and Koehler 1969 [38]. A number of recent variants of this idea go under the name *migration velocity analysis*, in which normal moveout correction is replaced by some variety of prestack migration (see for example Deregowski 1990 [9] or the volume Versteeg and Grau 1991 [45] for many examples and references). All use a measure of event coherence, either numerical (eg. stack power) or visual (flatness of events in velocity analysis or coherency panels, verticality of focusing spots) or both. These approaches are quite successful when their assumptions are justified, yet all are limited primarily by the need to search *systematically* a range of velocity models. The need for systematic search follows from the behaviour of the event coherence measures. Because of the extreme sensitivity of waveform data to velocities, the numerical measures are very nonconvex, and so must be sampled finely. The visual measures provide no means to improve the velocity automatically, so once again fine sampling is necessary. This systematic, finely sampled search perforce must be confined to a very small set of velocity models. For example one might consider piecewise linear interpolates of velocity profiles aligning events at a few control points. The possibility of unintended bias is obvious, and sometimes damaging. The synthetic model study Versteeg and Grau 1991 [45] provides compelling evidence of the limitations of these methods.

Another approach to velocity estimation is tomographic inversion or travel time analysis. The most successful instances allow the velocity model an *a priori* unlimited range of complexity, subject only to the constraint that the model explain the arrival times of events in time or (migrated) depth sections (Bishop et al. 1985 [2], Nolet 1987 [23], Scales 1987 [25], van Trier 1990 [42], Stork 1992 [30] and references cited there). The various measures of travel time error proposed in these works are quite well-behaved numerically - i.e. smooth and (weakly) convex over a wide range of velocity models - because they access travel time directly. Travel time is a quite tame function of velocities, in contrast to the waveform (amplitude) which is not, as mentioned earlier. The large degree of nonuniqueness inherent in the tomographic problem formulation may be controlled by imposition of geologically meaningful constraints (Stork and Clayton 1991 [31]).

A drawback of travel time inversion is that travel times are not the primary data of reflection seismology experiment, and must be picked from the waveform data. Travel time picking is most difficult in exactly those situations in which tomographic inversion is most needed: when the subsurface structure is complex (Ehinger 1992 [10], Harlan 1992 [15]). The possibility of bias also lurks in the multiplicity of events that are *not* picked in every tomographic exercise.

A third approach to velocity estimation is the simplest conceptually: introduce a complete physical model to predict the waveform data in detail. Following convention we will call this approach *inversion*. Most recent work has concentrated on the *best fit* formulation of inversion: choose a numerical measure of error between the data and the model-based synthetic, and adjust the model to minimize this measure. The most popular error measure is the mean square, because of its pleasant computational properties, though other measures have been used as well. One can find a comprehensive discussion of best fit inversion in Tarantola 1987 [41], including a specific formulation for inversion of reflection seismograms and an exposition of a philosophical justification of inversion based on Bayesian inference.

Differential semblance optimization is related to all three approaches to velocity estimation. It is a variant of least-squares inversion, as the objective function contains a mean square error term and so minimization yields an inversion. The most striking difference with ordinary least-squares inversion is that the model space is *expanded*, the model-parameters being regarded as *a priori experiment-dependent*. Thus each experiment (i.e. each shot, in the ordinary organization of seismic reflection data) is explained by an independent model. It turns out that each of the individual data-fitting problems is essentially quadratic (exactly quadratic, in the case of primaries-only modeling as in the sequel). Therefore production of an experiment dependent suite of models fitting independently the results of each experiment is a relatively straightforward task.

Of course such a multiplicity of models is infeasible, or incoherent: the earth is unique and all experiments should be explained by a single model. This observation motivates the addition of a *penalty* for variance amongst the experiment-dependent models to the objective function. This term plays the same role as semblance or stack power in migration velocity analysis, which measures the alignment of migrated images of the same depth point. However differential semblance optimization uses *models*, which should be *identical*, rather than migrated *images*, which must merely be aligned. Since models produced by inversion presumably do not contain the amplitude artifacts of migrated images, simple mean-square differences of neighboring models suffice as a reliable indicator of semblance.

Note that this conclusion requires both the use of *inverted models* and the application of a *differential* semblance measure to them, in contrast (and in par-

allel) to the use of *migrated images* and *integral or global* semblance measures in migration velocity analysis. Roelof Versteeg verified in his 1991 [44] thesis that mean-square nearest-neighbor differences do not give a robust velocity indicator if migrated images are used.

The separation into velocity and reflectivity submodels is a feature shared with travel time tomography. As observed by Stork and Clayton 1991 [31], it is essential in tomography to update both parts of the model simultaneously. The same is true of differential semblance optimization, which complicates the calculation of the objective gradient. The resemblance goes deeper than this, however. Analysis of the differential semblance Hessian shows that to lowest order the control over the velocity is exactly the same as that of a properly constructed tomographic Hessian (Song and Symes 1992 [29]). Moreover, the computed differential semblance velocity gradient shown in Figure 18 has exactly the appearance of a tomographic velocity gradient, with vertical smearing and side-lobes. Compare for example with the upper left-hand image in Figure 6 of Stork and Clayton 1992 [32], which has similar aperture and geometry.

The differential semblance criterion - that neighboring inverted reflectivities be the same - is the precise analog for full waveform models of the "flat image" or "minimum reflector variance" criterion used in migrated data domain tomography (van Trier 1990 [42], Stork 1992 [30], Harlan 1992 [15]).

In summary, differential semblance optimization appears to accomplish to some extent the goals of reflection traveltime tomography without event-picking, and of migration velocity analysis without requiring either sparsely parameterized models or interactive systematic search. It yields an infeasible-point approach to least squares inversion, and can be implemented efficiently with accurate gradient calculation.

Appendix A: Computation of the bilinear operator B

Computation of $B[q,\phi]$ is a straightforward application of the *adjoint state method*, as introduced in this context by Tarantola 1984 [39], Lailly 1983 [18], and others. The calculation described here is essentially identical to that given in Chavent and Jacewitz 1990 [6] for calculation of the gradient of a function of a migrated field with respect to velocity.

The field q here is a reflectivity, and will be used in place of r throughout the calculations. Recalling the wave equations 1 and 2, we see that the derivative $D_v F[v,q]\delta v$ of the forward map with respect to velocity is given by

$$D_v F[v,q]\delta v = \delta p(x_r, z_r, t; x_s)$$

where

$$\left(\frac{1}{v^2}\frac{\partial^2}{\partial t^2} - \nabla^2\right)\delta p_0 = 2\frac{\delta v}{v}\nabla^2 p_0$$

$$\left(\frac{1}{v^2}\frac{\partial^2}{\partial t^2} - \nabla^2\right)\delta p = 2\frac{\delta v}{v}\left[\nabla^2 p + 2q\nabla^2 p_0\right] + 2q\nabla^2\delta p_0$$

$$\delta p_0, \delta p \equiv 0, t < 0$$

We will use the "usual" adjoint field, i.e. the solution w of 9 for a test data section ϕ. Then:

$$< D_v F[v,q]\delta v, \phi >$$

$$= \int dx_s \int dx \int dz \int dt \delta p(x,z,t;x_s)\left(\int dx_r \phi(x_r,t;x_s)\delta(x-x_r, z-z_r)\right)$$

$$= \int dx_s \int dx \int dz \int dt \delta p\left(\frac{1}{v^2}\frac{\partial^2}{\partial t^2} - \nabla^2\right) w \qquad (15)$$

Because all fields obey Dirichlet boundary conditions at the surface and are of compact support in x for each t, integration by parts is at least formally permissible. We assume that f is in fact bandlimited; then all fields are in fact smooth away from the source point, and integration by parts is justified. In discretization, it must be correct that the discrete Laplace operator and boundary conditions and the quadrature rule used to replace the integrations allow exact calculation of adjoints. This is the case, for example, if one uses centered difference schemes, Dirichlet conditions on all boundaries implemented *via* the method of images, and the trapezoidal rule. Then the discrete operators are also self-adjoint, and summation by parts goes exactly as does the integration by parts which we present here.

Integration by parts and use of the equations satisfied by δp yields the continuation of 15:

$$= \int dx_s \int dx \int dz \int dt \left(2\frac{\delta v}{v} \left[\nabla^2 p + 2q\nabla^2 p_0 \right] + 2q\nabla^2 \delta p_0 \right) w$$

$$= \int dx \int dz \frac{2\delta v}{v} \int dx_s \int dt \left(\nabla^2 p + 2q\nabla^2 p_0 \right) w$$

$$+ 2 \int dx_s \int dx \int dz \int dt \delta p_0 \nabla^2 (qw) \qquad (16)$$

where in the last term we have performed another integration by parts.

To "get δv out" of the second term in the last field, we need to create a secondary backpropagated wavefield w_0 satisfying

$$\left(\frac{1}{v^2} \frac{\partial^2 w_0}{\partial t^2} - \nabla^2 w_0 \right) = 2\nabla^2 (qw)$$

$$w_0 \equiv 0, t > t_{max}$$

with homogeneous boundary conditions. Then substitution of the left-hand side of the w_0 equation in the second term of the second line in 16, integration by parts, and use of the equation satisfied by δp_0 gives the continuation of 16:

$$= \int dx \int dz \frac{2\delta v}{v} \int dx_s \int dt \left[\left(\nabla^2 p + 2q\nabla^2 p_0 \right) w + \left(\nabla^2 p_0 \right) w_0 \right]$$

$$= < \delta v, B[q, \phi] >$$

From this last equation we can immediately read off the prescription for computing $B[q, \phi]$ given in the text.

Appendix B: Derivation of the gradient correction

In this appendix we justify and derive the gradient correction presented in the text. The basis of this calculation is our use of Chebyshev iteration to produce the reflectivity estimate $r[v]_{est}$. We begin by writing the iteration in its normal three term recurrence form:

$$r_{k+1} = \omega_{k+1}(I - sN[v])r_k + (1 - \omega_{k+1})r_{k-1} + s\omega_{k+1}M[v]F_{data}$$

We write a prefix δ as usual for perturbation (directional derivative) with respect to δv of all of the quantities in this recursion. Note that the spectral bound of $N[v]$ is a stable function of v, since $N[v]$ is Hermitian, so we may regard the quantities ω_k and s as (locally) independent of v. We obtain

$$\begin{aligned} \delta r_{k+1} &= \omega_{k+1}(I - sN[v])\delta r_k + (1 - \omega_{k+1})\delta r_{k-1} \\ &\quad + s\omega_{k+1}(\delta M[v]F_{data} - \delta N[v]r_k), k = 1, 2, ... \end{aligned}$$

The key to understanding the relative importance of the pieces of this formula is that $M[v]$, as a migration operator, moves events around, whereas $N[v]$, as a simulation operator followed by a migration operator, is *kinematically neutral.* Thus $\delta M[v]$ involves infinitesimal phase shifts and so enhances high-frequency content, whereas $\delta N[v]$ does not. Accordingly we write

$$\delta r_k = y_k + z_k$$

where

$$\begin{aligned} y_{k+1} &= \omega_{k+1}(I - sN[v])y_k + (1 - \omega_{k+1})y_{k-1} \\ &\quad + s\omega_{k+1}(\delta M[v]F_{data} - \delta N[v]r_{k_{max}}), k = 1, 2, ... \end{aligned}$$

$$\begin{aligned} z_{k+1} &= \omega_{k+1}(I - sN[v])z_k + (1 - \omega_{k+1})z_{k-1} \\ &\quad + s\omega_{k+1}(\delta N[v](r_{k_{max}} - r_k)), k = 1, 2, ... \end{aligned}$$

The inhomogeneous term in the recursion for y_k is *independent of* k, so y_k is the result of the application of the same polynomial as is r_k:

$$r_k = P_k(N[v])M[v]F_{data}$$

$$\begin{aligned} y_k &= P_k(N[v])(\delta M[v]F_{data} \\ &- \delta N[v]r_{k_{max}}) \end{aligned}$$

Otherwise put, we can compute y_k by *re-running the iteration which produced* r_k, *with different data.* Moreover the data for y_k contains $\delta M[v]F_{data}$, which

is the source of all instability in δr_k. The data for the z_k recursion, on the other hand, is k-dependent, which renders z_k essentially uncomputable from the point of view of the adjoint calculations to come. However the data for z_k involves only $\delta N[v]$, which is tame, and $r_{k_{max}} - r_k$ which tends to become small as $k \rightarrow k_{max}$.

Therefore we throw z_k away and replace δr_k by y_k in the second term of 10, which also involves the normal residual e_k. So after k iterations we actually have

$$D\tilde{J}_k[v]\delta v =< D_v F[v, r_k]\delta v, F[v, r_k] - F_{data} > - < y_k, e_k >$$

Now

$$\begin{aligned}
< y_k, e_k > &= < P_k(N[v])(\delta M[v]F_{data} - \delta N[v]r_{k_{max}}), e_k > \\
&= < \delta M[v]F_{data} - \delta M[v]F[v, r_{k_{max}}] \\
&\quad -M[v]D_v F[v, r_{k_{max}}]\delta v, P_k(N[v])e_k > \\
&= < F_{data} - F[v, r_{k_{max}}], D_v F[v, P_k(N[v])e_k]\delta v > \\
&\quad - < D_v F[v, r_{k_{max}}]\delta v, F[v, P_k(N[v])e_k] > \\
&= < \delta v, B[P_k(N[v])e_k, F_{data} - F[v, r_{k_{max}}]] \\
&\quad -B[r_{k_{max}}, F[v, P_k(N[v])e_k] >
\end{aligned}$$

Here we have used the definition of $N[v]$, the adjoint relation

$$< F[v, q], \phi >=< q, M[v]\phi >$$

and relations between derivatives that follow from these.

Adding the naive term expressed in terms of B as before and setting $k = k_{max}$ we get for gradient in the sense of the Sobolev s-norm

$$\begin{aligned}
grad\tilde{J}[v] &= \Lambda^{-2s}(B[r_{k_{max}} - P_{k_{max}}(N[v])e_{k_{max}}, F[v, r_{k_{max}}] - F_{data}] \\
&\quad -B[r_{k_{max}}, F[v, P_{k_{max}}(N[v])e_{k_{max}}]
\end{aligned}$$

It is worth listing just what is involved in this formula:

1. one round of Chebyshev iteration to compute the reflectivity

   $$r_{k_{max}} = P_{k_{max}}(N[v])M[v]F_{data}$$

   the residual $F[v, r_{k_{max}}] - F_{data}$ and the normal residual $e_{k_{max}}$;

2. another round of Chebyshev iteration to compute field

   $$P_{k_{max}}(N[v])e_{k_{max}};$$

3. two generalized adjoint state calculations to compute the values of B.

References

[1] G. BEYLKIN. Imaging of discontinuities in the inverse scattering problem by inversion of a causal generalized radon transform. *J. Math. Phys.*, 26:99–108, 1985.

[2] T.N. BISHOP, K.P. BUBE, R.T. CUTLER, R.T. LAINGAN, P.L. LOVE, J.R. RESNICK, R.T. SHUEY, D.A. SPINDLER, and H.W. WYLD. Tomographic determination of velocity and depth in laterally varying media. *Geophysics*, 50:903–923, 1987.

[3] N. BLEISTEIN. On the imaging of reflectors in the earth. *Geophysics*, 52(7):931–942, 1987.

[4] A. BOURGEOIS, B.F. JIANG, and P. LAILLY. Linearized inversion: A significant step beyond pre-stack migration. *Geophysics J. Int.*, 99:435–445, 1989.

[5] D. CAO, S. SINGH, and A. TARANTOLA. Simultaneous inversion for background velocity and impedance maps. *Geophysics*, 55:458–469, 1990.

[6] G. CHAVENT and C. JACEWITZ. Automatic determination of background velocities by multiple migration fitting. In *Proc. 60th Annual Meeting*, pages 1263–1266, San Francisco, 1990. Society of Exploration Geophysicists. Expanded abstract.

[7] J.F. CLAERBOUT. *Imaging the Earth's Interior.* Blackwell Scientific Publishers, Oxford, 1985.

[8] M.A. DABLAIN. The application of high-order differencing to the scalar wave equation. *Geophysics*, 51:54–66, 1986.

[9] S.M. DEREGOWSKI. Common offset migrations and velocity analysis. *First Break*, 8:225–234, 1990.

[10] A. EHINGER. How to pick traveltime in complex structures? In *54th Meeting.* European Association of Exploration Geophysicists, 1992. Abstract.

[11] B. ENGQUIST and A. MAJDA. Absorbing boundary conditions for the numerical simulation of waves. *Math. Comp.*, 31:629–651, 1977.

[12] R. FLETCHER. *Practical methods of optimization, I:Unconstrained Optimization.* Wiley & Sons, New York, 1980.

[13] F.G. FRIEDLANDER. *Sound Pulses.* Cambridge University Press, 1958.

[14] O. GAUTHIER, A. TARANTOLA, and J. VIRIEUX. Two-dimensional nonlinear inversion of seismic waveforms. *Geophysics*, 51:1387–1403, 1986.

[15] W. HARLAN. Robust strategies for reflection and crosswell traveltime tomography. In *54th Meeting*. European Association of Exploration Geophysicists, 1992. Abstract.

[16] J. HELGESEN and P. KOLB. Inversion of 1-D acoustic medium from real seismic data. In *Proc. 59th Annual International Meeting*, pages 1000–1003. Society of Exploration Geophysicists, 1989. Expanded abstract.

[17] P. KOLB, F. COLLINO, and P. LAILLY. Prestack inversion of a 1d medium. *Proceedings of IEEE 74*, pages 498–506, 1986.

[18] P. LAILLY. The seismic inverse problem as a sequence of before-stack migrations. In J.B. Bednar et al., editors, *Conference on Inverse Scattering: Theory and Applications*, pages 206–220. SIAM, Philadelphia, 1983.

[19] A.R. LEVANDER. Fourth order finite difference P-SV seismograms. *Geophysics*, 53:1425–1434, 1988.

[20] R.M. LEWIS. *Source-Velocity Identification for a Layered Model of Reflection Seismology*. PhD thesis, Department of Mathematical Sciences, Rice University, Houston, Texas, U.S.A, 1989.

[21] K. MOSEGARD and A. TARANTOLA. Monte Carlo analysis of geophysical inverse problems. In *Proc. 61st Annual International Meeting and Exposition*. Society of Exploration Geophysicists, 1991.

[22] NOBLE. PhD thesis, Université de Paris VI, 1992.

[23] G. NOLET. *Seismic Tomography*. D. Reidel & Co., Dordrecht, 1987.

[24] F. SANTOSA and W. SYMES. An analysis of least-squares velocity inversion. Geophysical Monograph 4, Soc. of Expl. Geophys., Tulsa, 1989.

[25] J.A. SCALES. Tomographic inversion *via* the conjugate gradient method. *Geophysics*, 52:179–185, 1987.

[26] J.A. SCALES, M.L. SMITH, and T.L. FISCHER. Global optimization methods for highly nonlinear inverse problems. In G. Cohen, L. Halpern, and P. Joly, editors, *Mathematical and Numerical Aspects of Wave Propagation Phenomena*, pages 434–444. SIAM, Philadelphia, 1991.

[27] M.K. SEN and P. STOFFA. Nonlinear multiparameter optimization using genetic algorithms: Inversion of plane wave seismograms. *Geophysics*, 56:1794–1810, 1991.

[28] M.K. SEN and P. STOFFA. Nonlinear one-dimensional seismic waveform inversion using simulated annealing. *Geophysics*, 56:1624–1636, 1991.

[29] H. SONG and W. SYMES. A differential semblance formulation of crosswell tomography. Technical report, Department of Computational and Applied Mathematics, Rice University, Houston, TX, 1992.

[30] C. STORK. Reflection tomography in the postmigrated domain. *Geophysics*, 57:680–692, 1992.

[31] C. STORK and R. CLAYTON. Linear aspects of tomographic velocity analysis. *Geophysics*, 56:483–495, 1991.

[32] C. STORK and R. CLAYTON. Using constraints to address the uncertainties of automated prestack velocity analysis. *Geophysics*, 57:404–419, 1992.

[33] W. SYMES. A differential semblance algorithm for the inverse problem of reflection seismology. *Computers and Math. with Appl.*, 22(4/5):147–178, 1991.

[34] W. SYMES. Layered velocity inversion: a model program from reflection seismology. *SIAM J. Math. Anal.*, 22:680–716, 1991.

[35] W. SYMES. Non-interactive estimation of the Marmousi velocity model by differential semblance optimization: Initial trials. In G. Grau and R. Versteeg, editors, *The Marmousi Experience: Proceedings of the EAEG Workshop on Practical Aspects of Inversion*, The Hague, 1991.

[36] W. SYMES. A differential semblance criterion for inversion of multioffset seismic reflection data. *J. Geophys. Res.*, 1992. to appear.

[37] W. SYMES and J. CARAZZONE. Velocity inversion by differential semblance optimization. *Geophysics*, 56(5):654–663, 1991.

[38] M.T. TANER and F. KOEHLER. Velocity spectra: digital computer derivation and application of velocity functions. *Geophysics*, 34:859–881, 1969.

[39] A. TARANTOLA. Inversion of seismic reflection data in the acoustic approximation. *Geophysics*, 49:1259–1266, 1984.

[40] A. TARANTOLA. A strategy for nonlinear elastic inversion of seismic reflection data. *Geophysics*, 51:1893–1903, 1986.

[41] A. TARANTOLA. *Inverse Problem Theory*. Elsevier, 1987.

[42] J. VAN TRIER. PhD thesis, Stanford University, 1990.

[43] R. VARGA. *Matrix Iterative Analysis*. Prentice-Hall, Englewood Cliffs, NJ, 1962.

[44] R. VERSTEEG. PhD thesis, Institut Francais du Petrole, 1991.

[45] R. VERSTEEG and G. GRAU. Practical aspects of inversion: The Marmousi experience. Proceedings of the EAEG, The Hague, 1991.

Table 1
Performance of NLCG over 8 Iterations

i	J_i	J_i/J_0	$\|\|gradJ_i\|\|$	$\|\|gradJ_i\|\|/\|\|gradJ_0\|\|$
0	0.2394E+01	0.1000E+01	0.7779E+00	0.1000E+01
1	0.1337E+01	0.5587E+00	0.7746E-01	0.9958E-01
2	0.1185E+01	0.4952E+00	0.3577E+00	0.4599E+00
3	0.8021E+00	0.3351E+00	0.2324E+00	0.2987E+00
4	0.6488E+00	0.2710E+00	0.2674E+00	0.3437E+00
5	0.6178E+00	0.2581E+00	0.2176E-01	0.2797E-01
6	0.5958E+00	0.2488E+00	0.5201E-01	0.6686E-01
7	0.5736E+00	0.2396E+00	0.1981E-01	0.2546E-01
8	0.5734E+00	0.2394E+00	0.9622E-02	0.1237E-01

Table 2
Assessment of Computed Gradient Accuracy

Normal residual	0.09	0.03
Finite difference	-45.12	-48.59
Computed derivative, uncorrected gradient	-42.89	-46.47
Difference in %	4.9	4.4
Computed derivative, corrected gradient	-46.19	-48.74
Difference in %	2.4	0.3

Figure Captions - GP paper

1. A cartoon illustrating the smoothness of differential semblance as a function of velocity. The cube represents the prestack inverted reflectivity, plotted as a function of depth, (global) horizontal coordinate, and shot location. We imagine that the (linearized) inversion has been performed with substantially incorrect velocity in Figure 1a and with the correct velocity in Figure 1b. Accordingly the common depth point gathers (often called coherency or common image panels) display noticeable moveout for the incorrect velocity (a), but consist of identical traces for the correct velocity (b). In reality edge effects will prevent complete short-independence in the correct velocity case. The *relative* moveout between shot gathers (or equally well between traces on a common depth point gather) is small, amounting to much less than a wavelength, because the shot sampling is fine. Therefore a small change in velocity will result in a small change in the differences between neighboring traces on the common depth point gather, therefore a small change in the differential semblance as defined in the text. The stack power, on the other hand, may change dramatically, because it involves the interaction of distant traces which may be entirely out of phase.

2. The velocity (a) and reflectivity (b) profiles v_{exact} and r_{exact} used in the numerical experiments to generate the layered model. Horizontal scale is meters, and vertical scales are meters per millisecond and dimensionless, respectively.

3. The two-shot data set generated from the layered model in Figure 2. The data has been muted linearly to remove the water-bottom event, which becomes contaminated with a perturbational head-wave onset at larger offsets. Vertical scale is milliseconds. Trace spacingis 50 m, with near offset of 150 m and far offset of 1800 m. Shot spacing is 96 m.

4. Samples of the differential semblance objective function along the line segment between the constant velocity $v_{base} \equiv 1.5m/ms$ and v_{exact}. The segment is parameterized by

$$h \mapsto v_{base} + h * [v_{exact} - v_{base}]$$

so that $h = 0$ is constant velocity and $h = 1$ is the target v_{exact}. The curves correspond to $\sigma = 1.0$ (dots), 10.0 (dash-dots), 100.0 (dashes), and 1000.0 (solid line). Note that the minimizer is located in all cases at or near $h = 1.0$, and that the curves appear to be converging as $\sigma \to \infty$. This convergence effect is peculiar to 2-shot data sets, as pointed our in the text.

5. The result of four nonlinear conjugate gradient steps (solid) versus v_{exact} (dashes). Note that the overshoot from the surface to roughly 400 m is balanced by undershoot between 400 and 700 m. Relatively little reflected energy is present in the data (Figure 3) between the surface and 700 m. Scales same as Figure 2a.

6a. The reflectivity gather from shot 1, corresponding to the velocity inversion in Figure 5. Note that, apart from edge diffractions, the gather is quite flat. Vertical scale is meters, trace spacing is 16 m.

b. The reflectivity gather obtained by linearized inversion of the data from Figure 3 at the target velocity v_{exact}. Visually, flatness roughly equivalent to that displayed in Figure 6a. Inversion performed with the same parameters as in Figure 6a.

c. The reflectivity gather obtained by linearized inversion of the data from Figure 3 at the constant velocity v_{base}. Note the marked moveout. Since the data is translation invariant, this single shot gather is actually equivalent to a common depth gather (common. image panel).

7. Data predicted from the inverted model (Figures 5, 6a).

8. Error (difference between Figures 3 and 7), plotted on same amplitude scale as Figure 7. Error level is approximately 15% RMS.

9. Result of eight nonlinear conjugate gradient steps (solid line), compared with the result of four steps (dashes) and target (dash-dots).

10. Reflectivity corresponding to the velocity inversion in Figure 9.

11. Velocities obtained after four steps of nonlinear conjugate gradient iteration using smoothing lengths of 300 m (dash-dots), 400m (solid), 500 m (dashes), and target (dots).

12a. Reflectivity gather produced by inversion at the 300 m velocity model (Figure 11, dash-dots).

b. Reflectivity gather produced by inversion at the 400 m velocity model (Figure 11, solid).

c. Reflectivity gather produced by inversion at the 500 m velocity model (Figure 11, dashes).

13. Grey-scale plot of DSO gradient at constant velocity, computed for a 35-shot data set generated from the layered model (Figure 2). Black is negative, white is positive. Thus this gradient suggests a general increase in velocity, essentially layered within the inversion aperture. Edge anomalies occur in the vicinities of the first and last shots in the line.

14. Several vertical cross sections of the gradient in Figure 13, illustrating the extent to which it is layered.

15. Grey-scale plot of layered model with high velocity near surface anomaly.

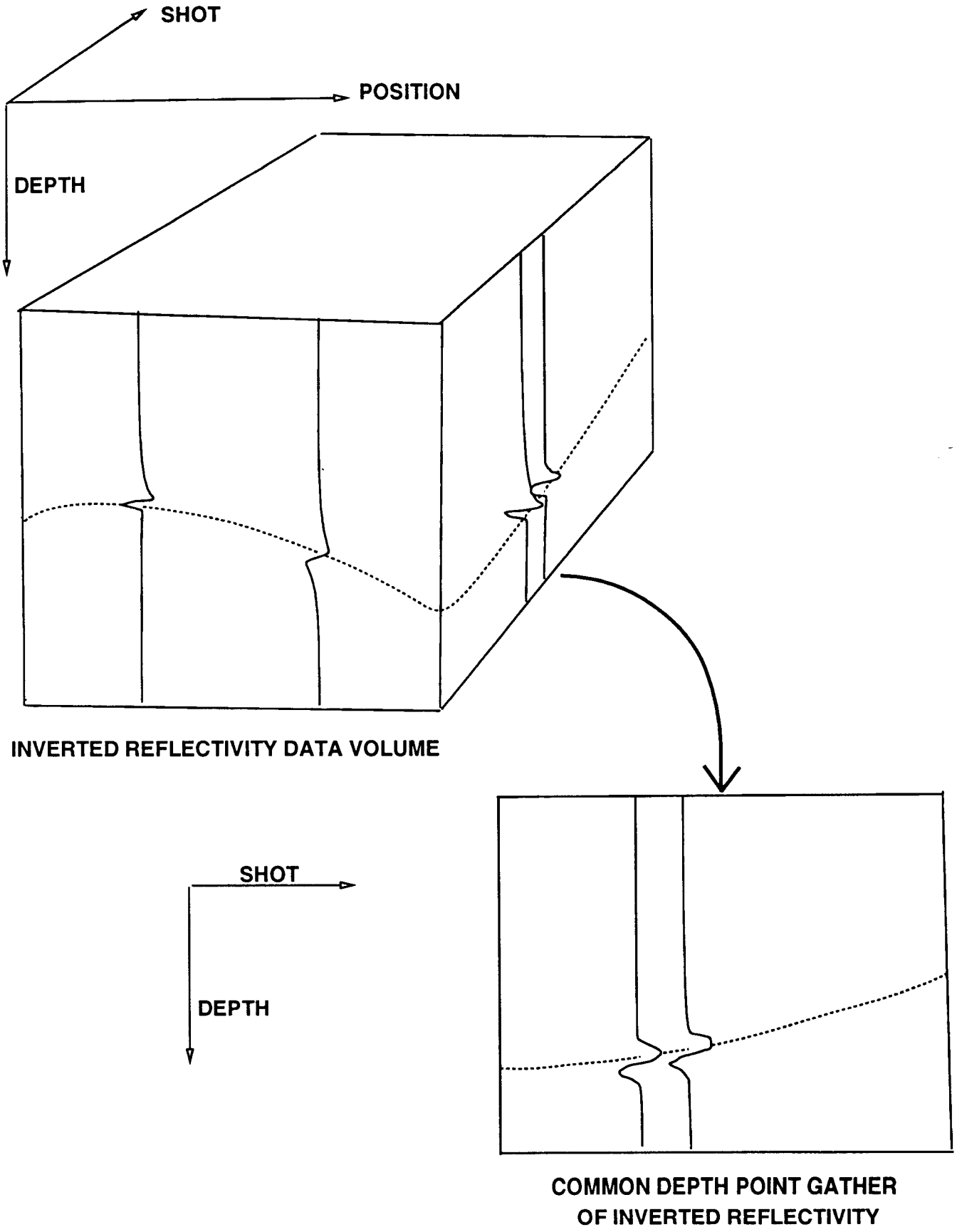
16. Some gathers from the 35-shot data set derived from the anomaly model of Figure 13 and the layered reflectivity of Figure 2b.

17. Gradient at constant background velocity. Evidence of nonlayering present, but no clear indication of its form.

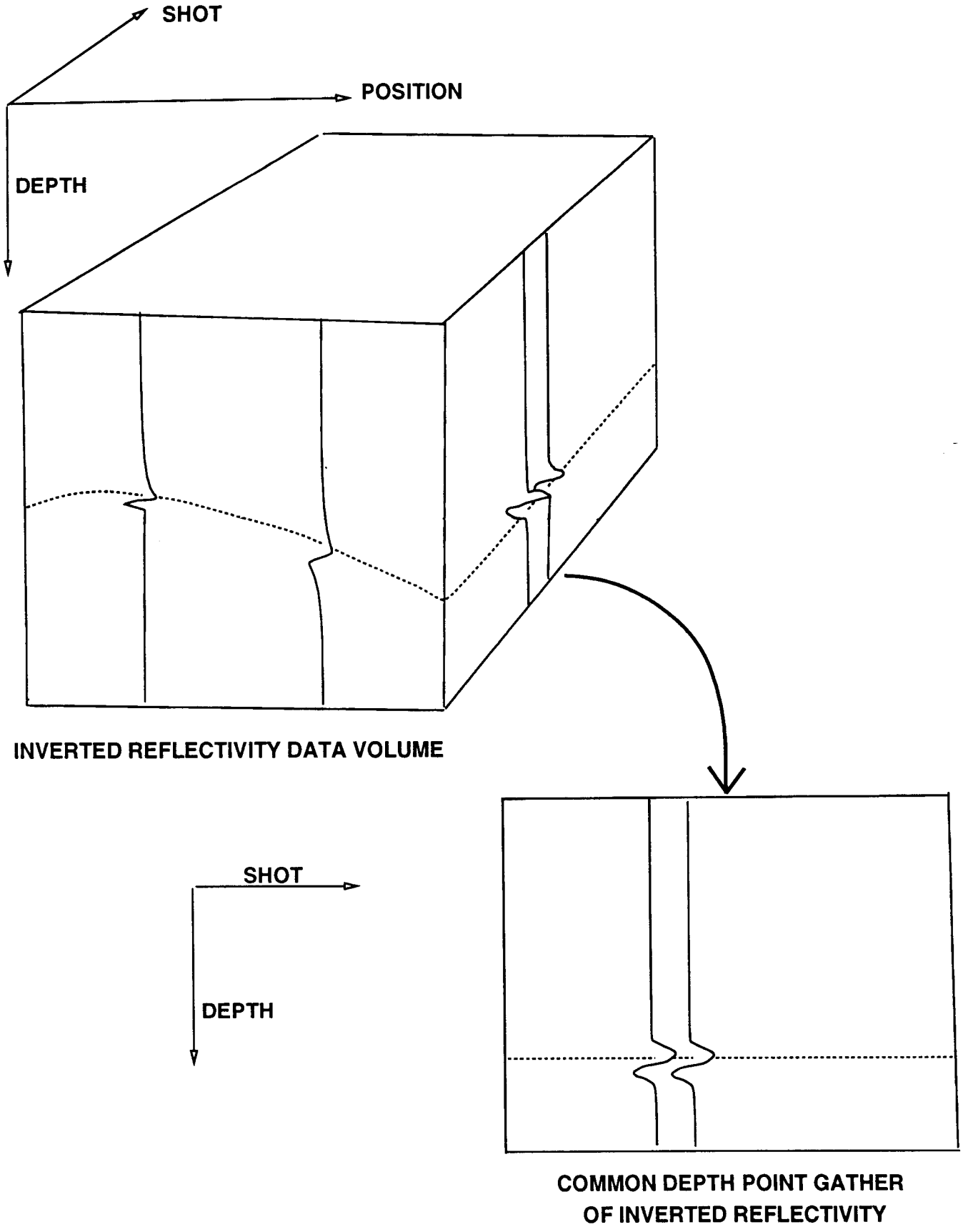
18. Gradient at layered background (= v_{exact} shown in Figure 2, i.e. without the anomaly). Location and type of the anomaly clearly defined. The vertical smearing and side lobes are characteristic of all limited-aperture tomographic reconstruction methods.

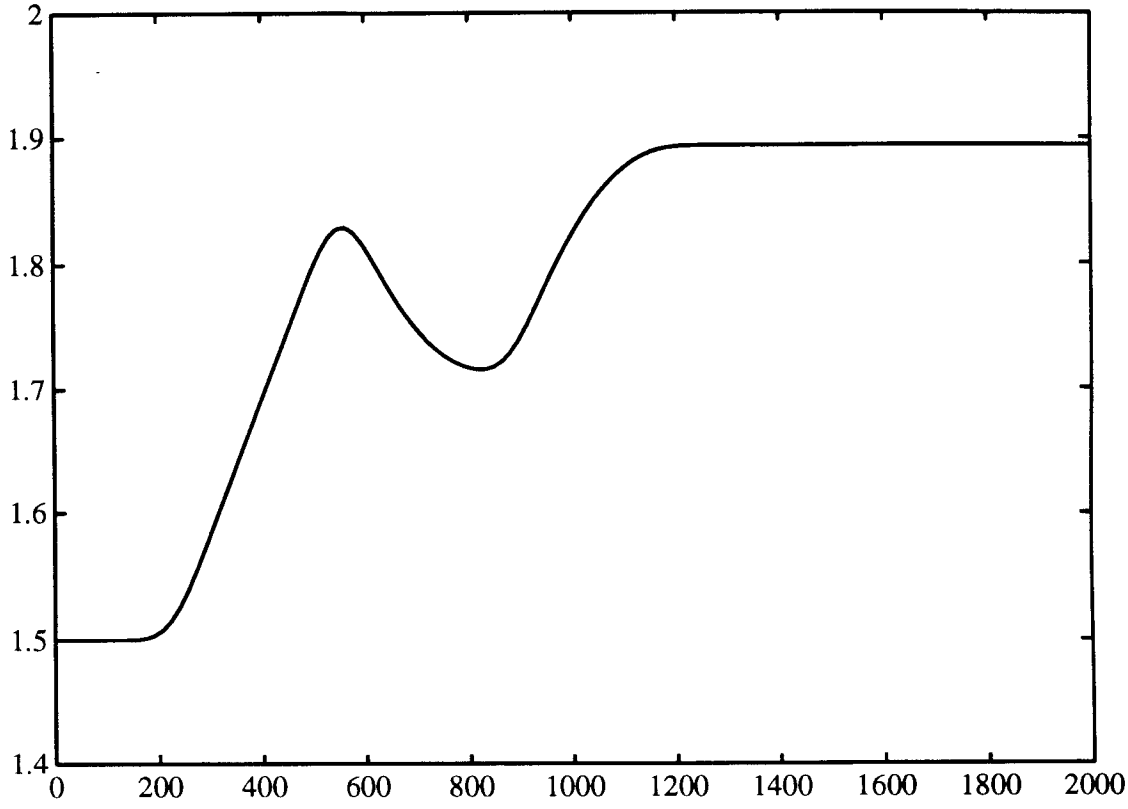
19. Horizontal cross section through layered background gradient (Figure 18) at 240 m depth.

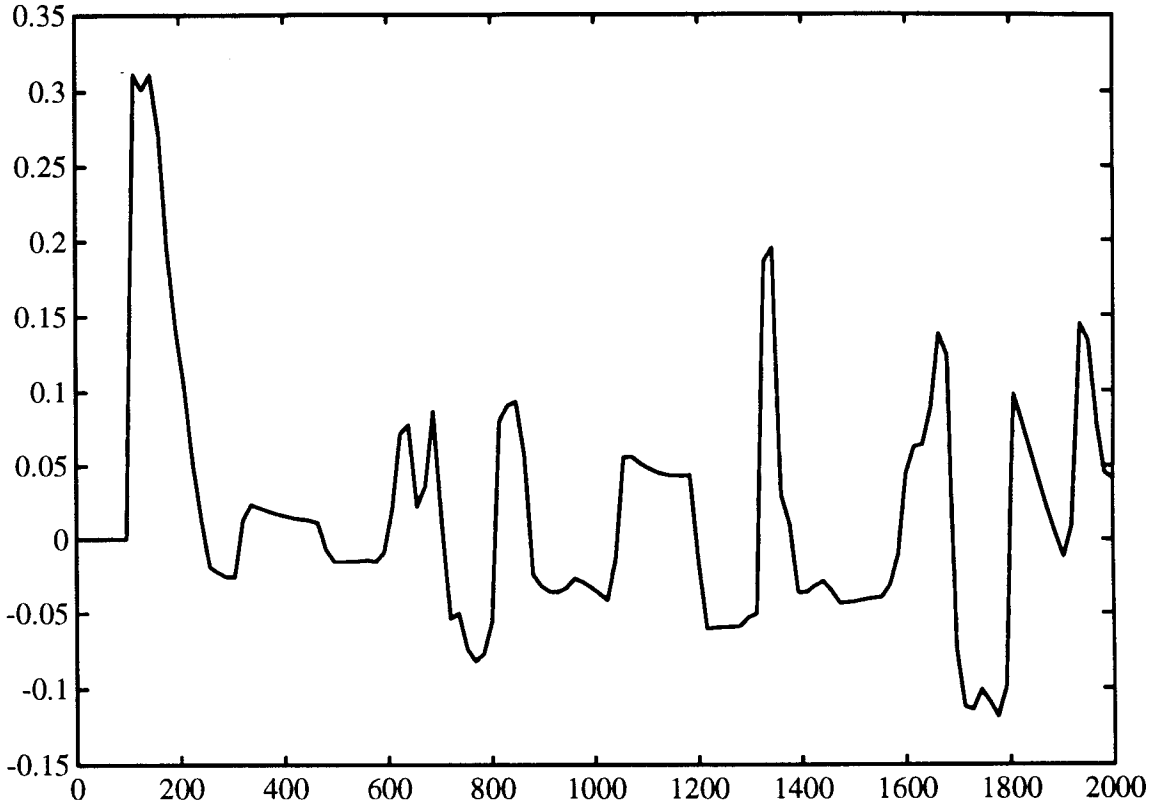
1(a) REFLECTIVITY INVERSION AT INCORRECT VELOCITY

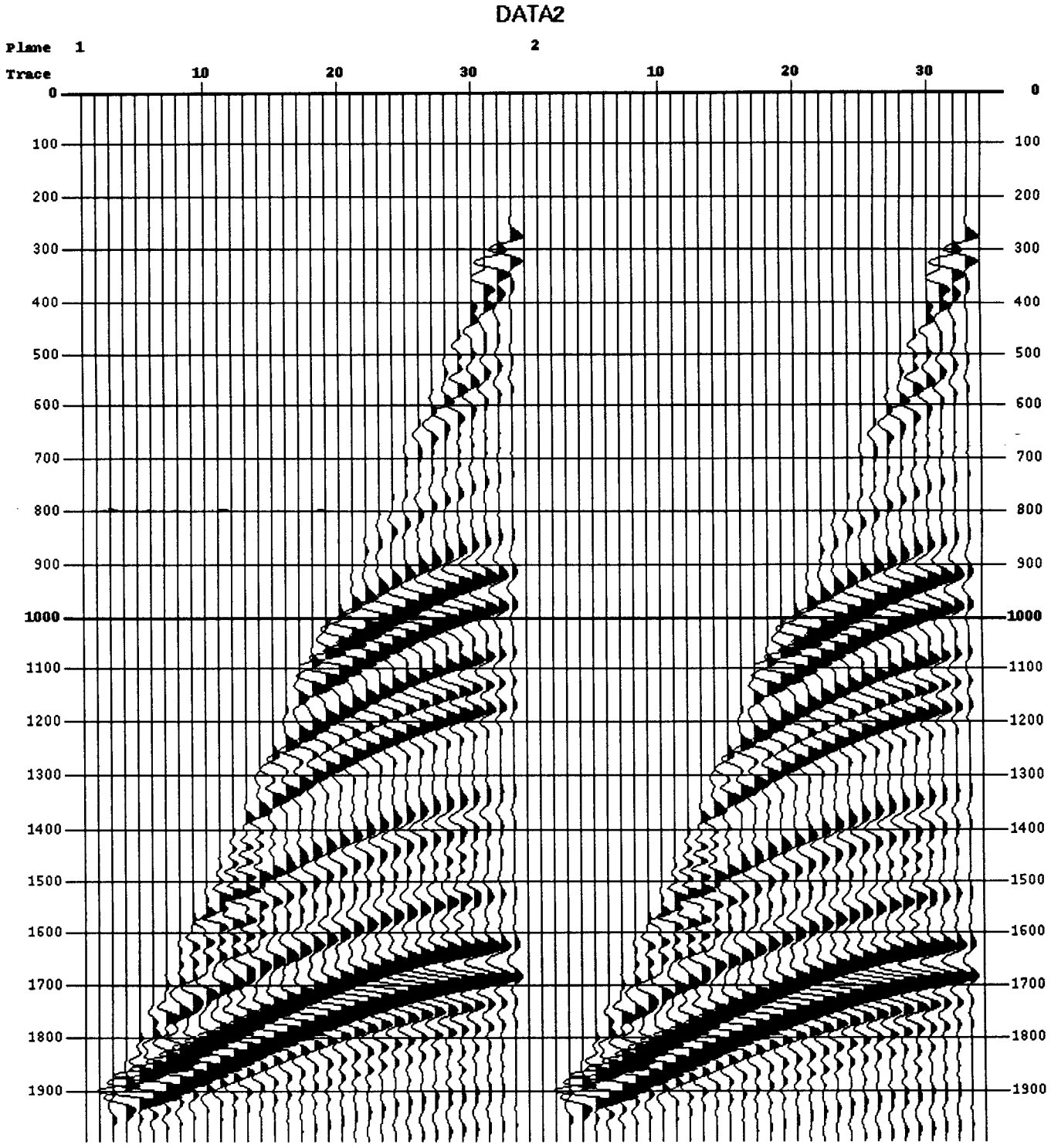


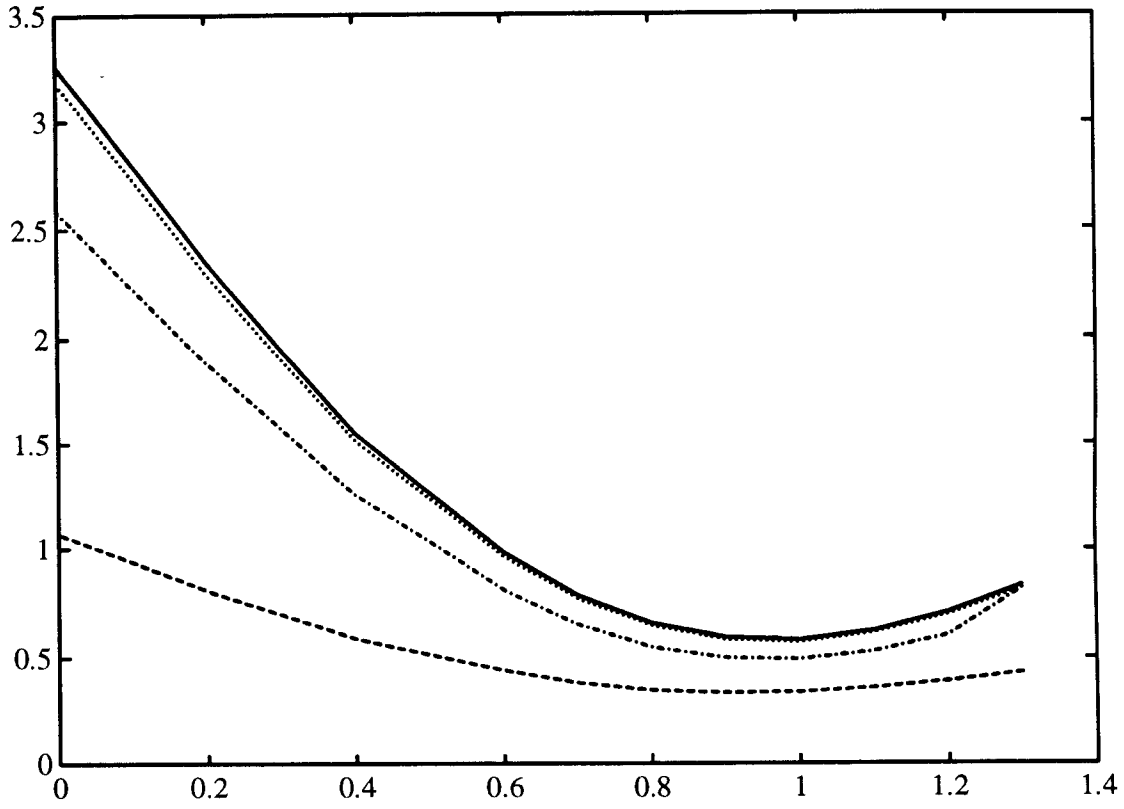
1(b) REFLECTIVITY INVERSION AT CORRECT VELOCITY

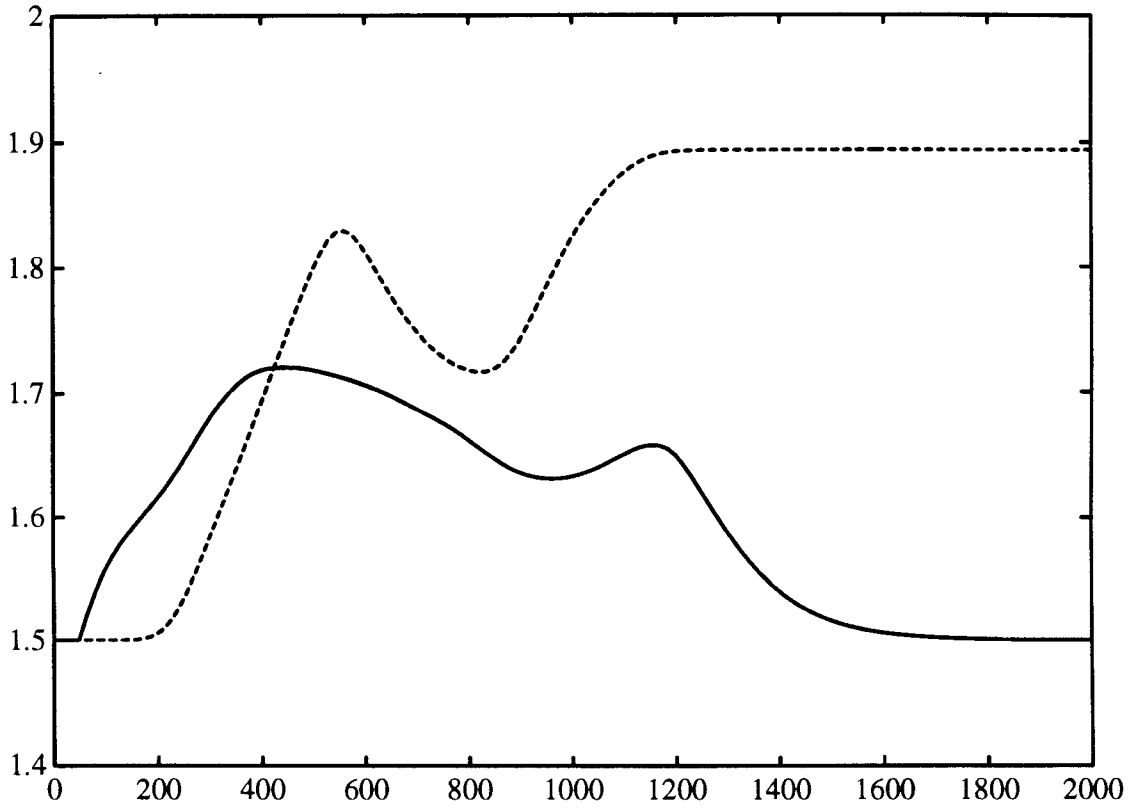


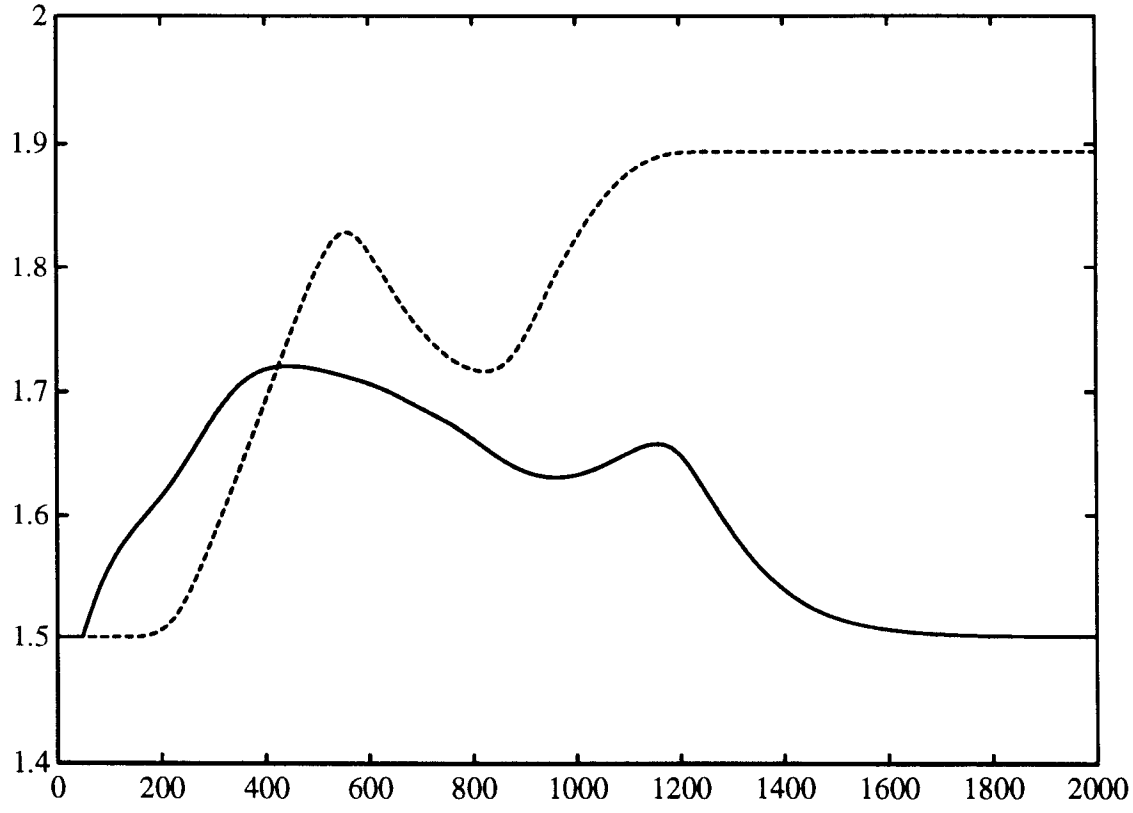

2
1.9
1.8
1.7
1.6
1.5
1.4
0
200
400
600
800
1000
1200
1400
1600
1800
2000

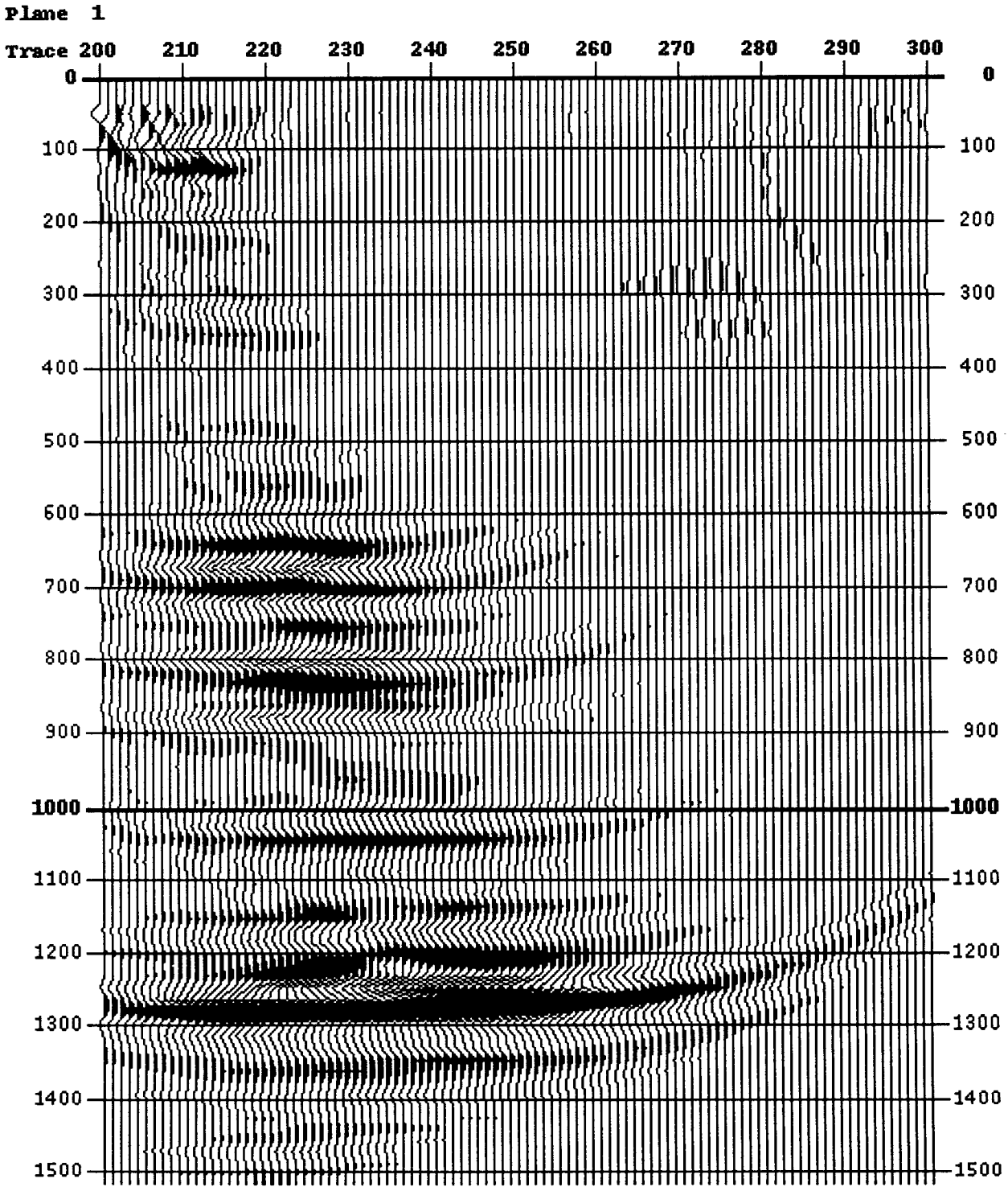

0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-0.05
-0.1
-0.15
0
200
400
600
800
1000
1200
1400
1600
1800
2000


DATA2
Plane 1
2
Trace
10
20
30
10
20
30
0
100
200
300
400
500
600
700
800
900
1000
1100
1200
1300
1400
1500
1600
1700
1800
1900

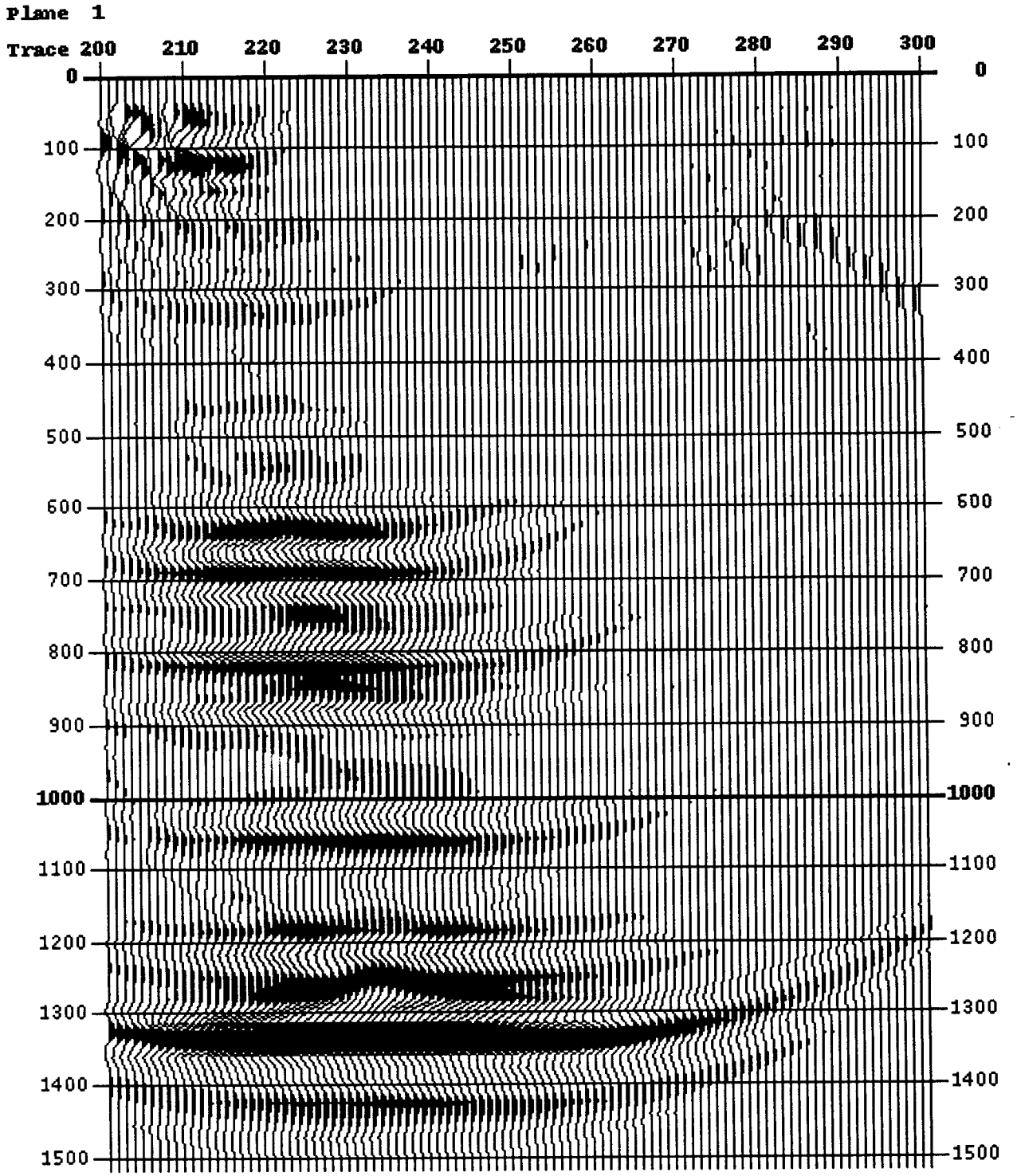

3.5
3
2.5
2
1.5
1
0.5
0
0
0.2
0.4
0.6
0.8
1
1.2
1.4

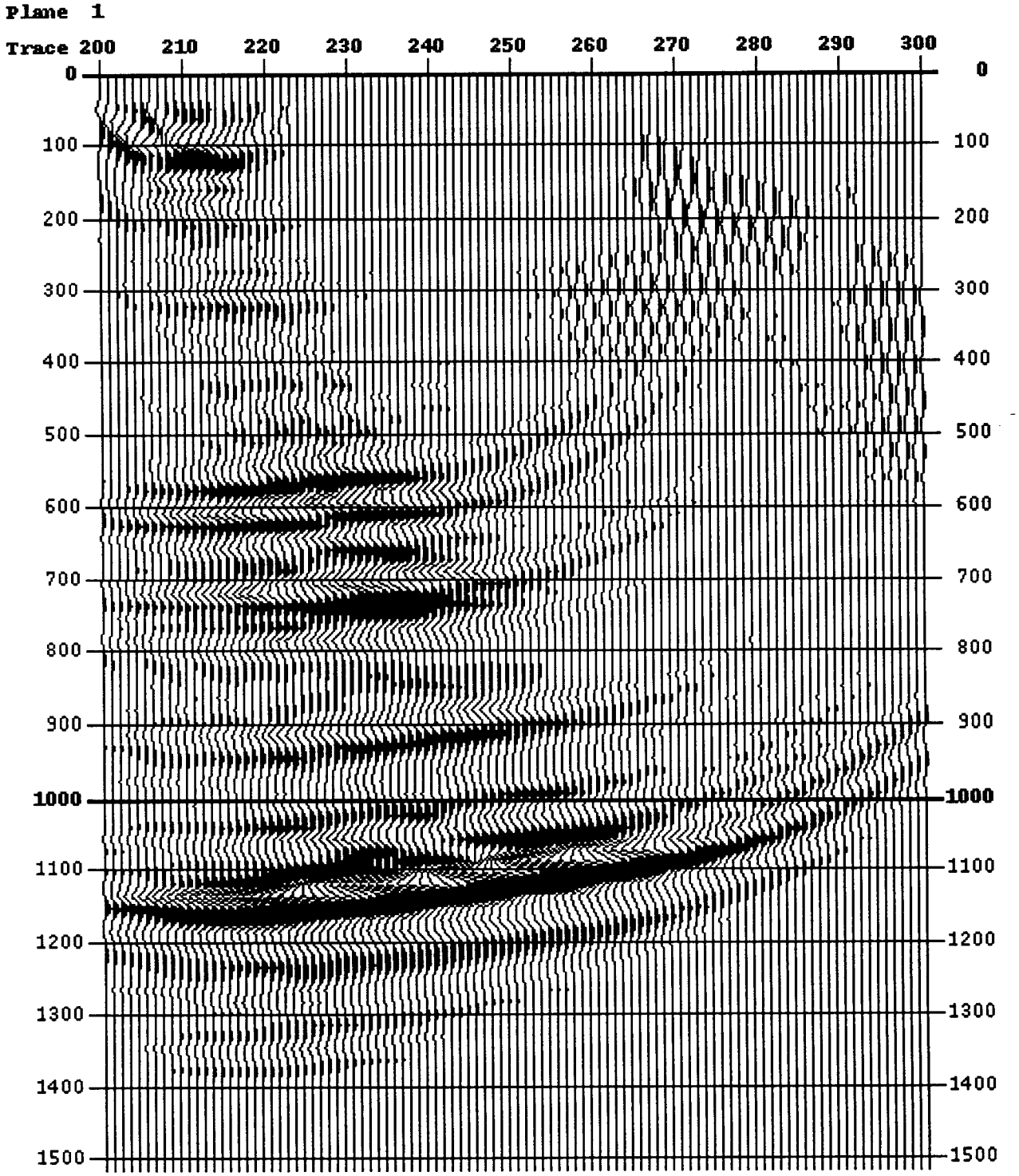

2
1.9
1.8
1.7
1.6
1.5
1.4
0
200
400
600
800
1000
1200
1400
1600
1800
2000


2
1.9
1.8
1.7
1.6
1.5
1.4
0
200
400
600
800
1000
1200
1400
1600
1800
2000

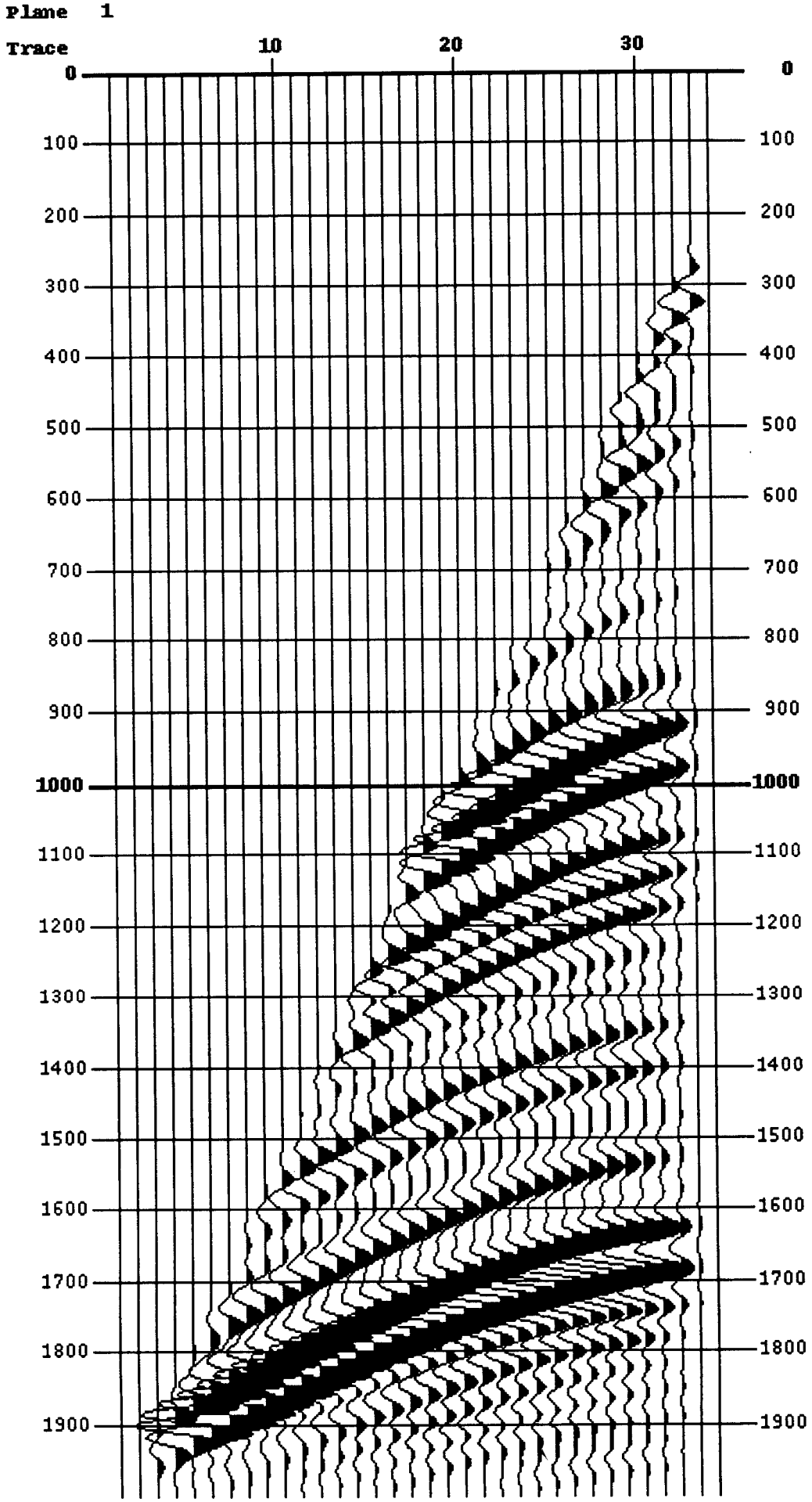

Plane 1
Trace 200 210 220 230 240 250 260 270 280 290 300
0 100 200 300 400 500 600 700 800 900 1000 1100 1200 1300 1400 1500

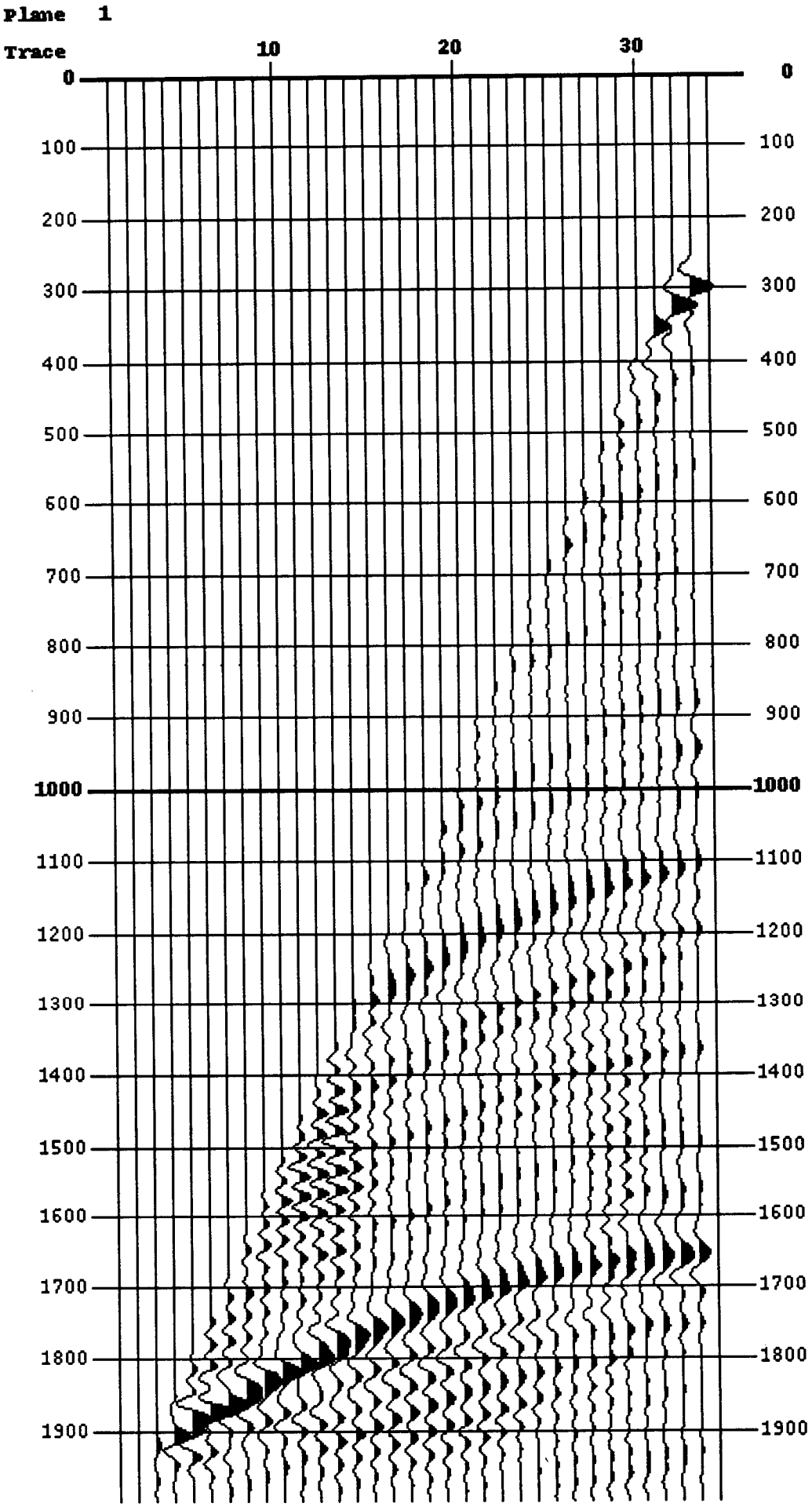
6a

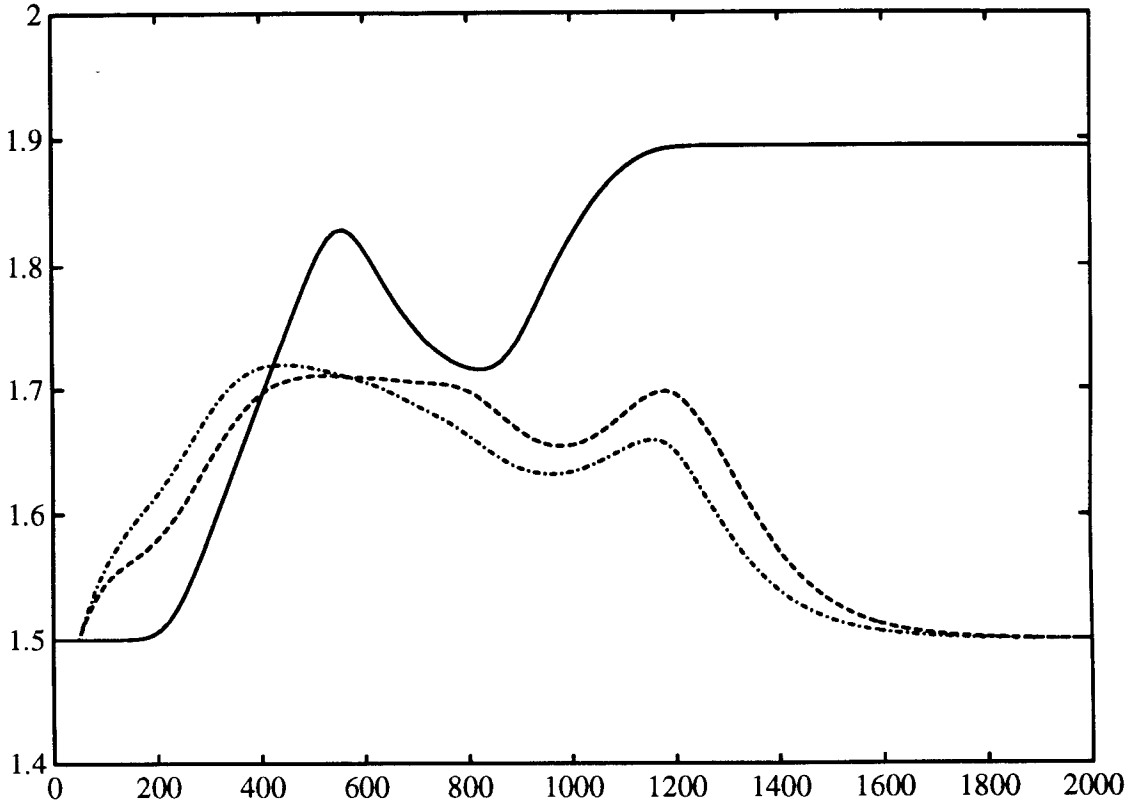

Plane 1
Trace 200
210
220
230
240
250
260
270
280
290
300
0
100
200
300
400
500
600
700
800
900
1000
1100
1200
1300
1400
1500

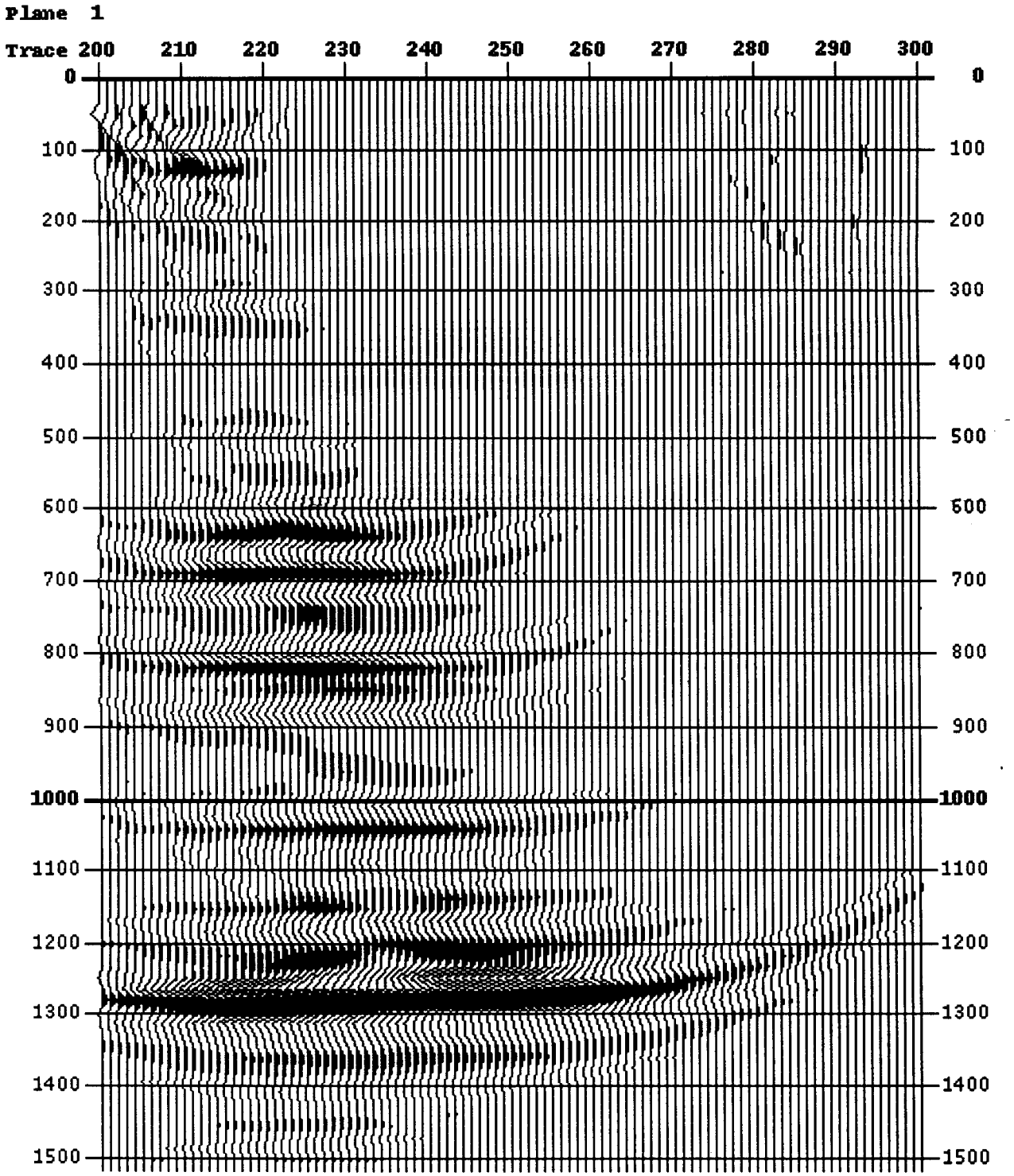

Plane 1
Trace 200 210 220 230 240 250 260 270 280 290 300
0 100 200 300 400 500 600 700 800 900 1000 1100 1200 1300 1400 1500

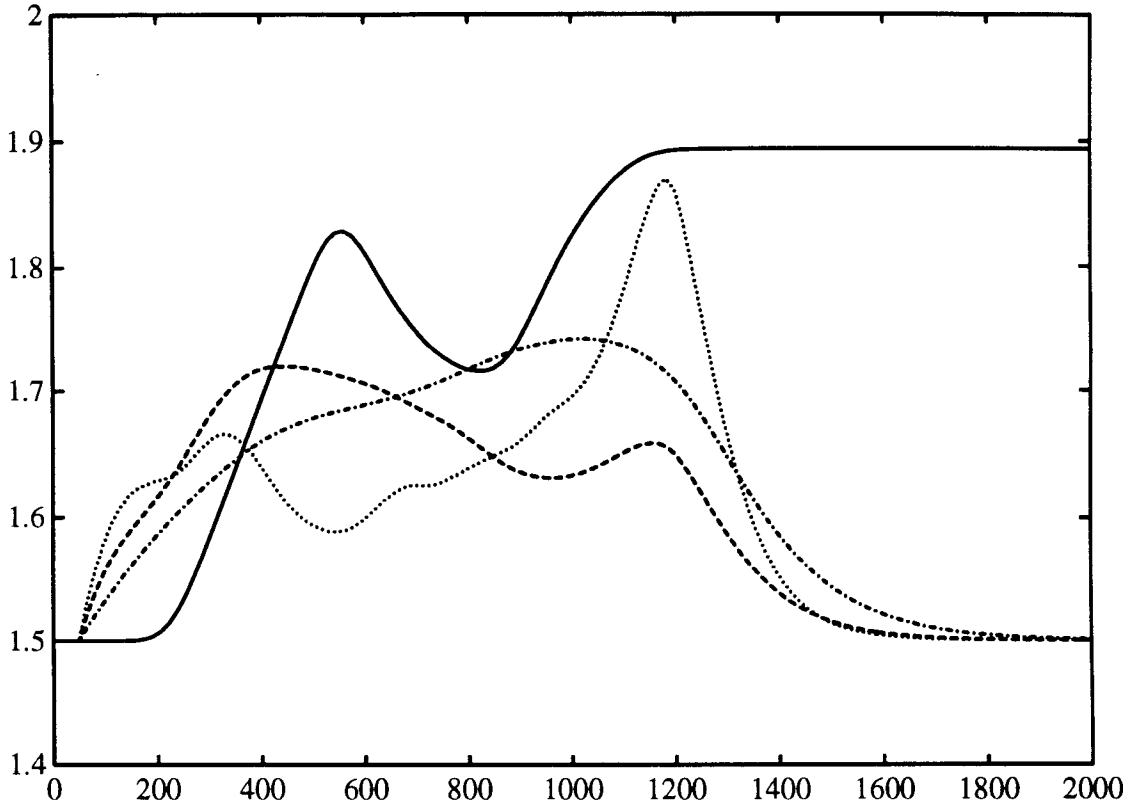
6c

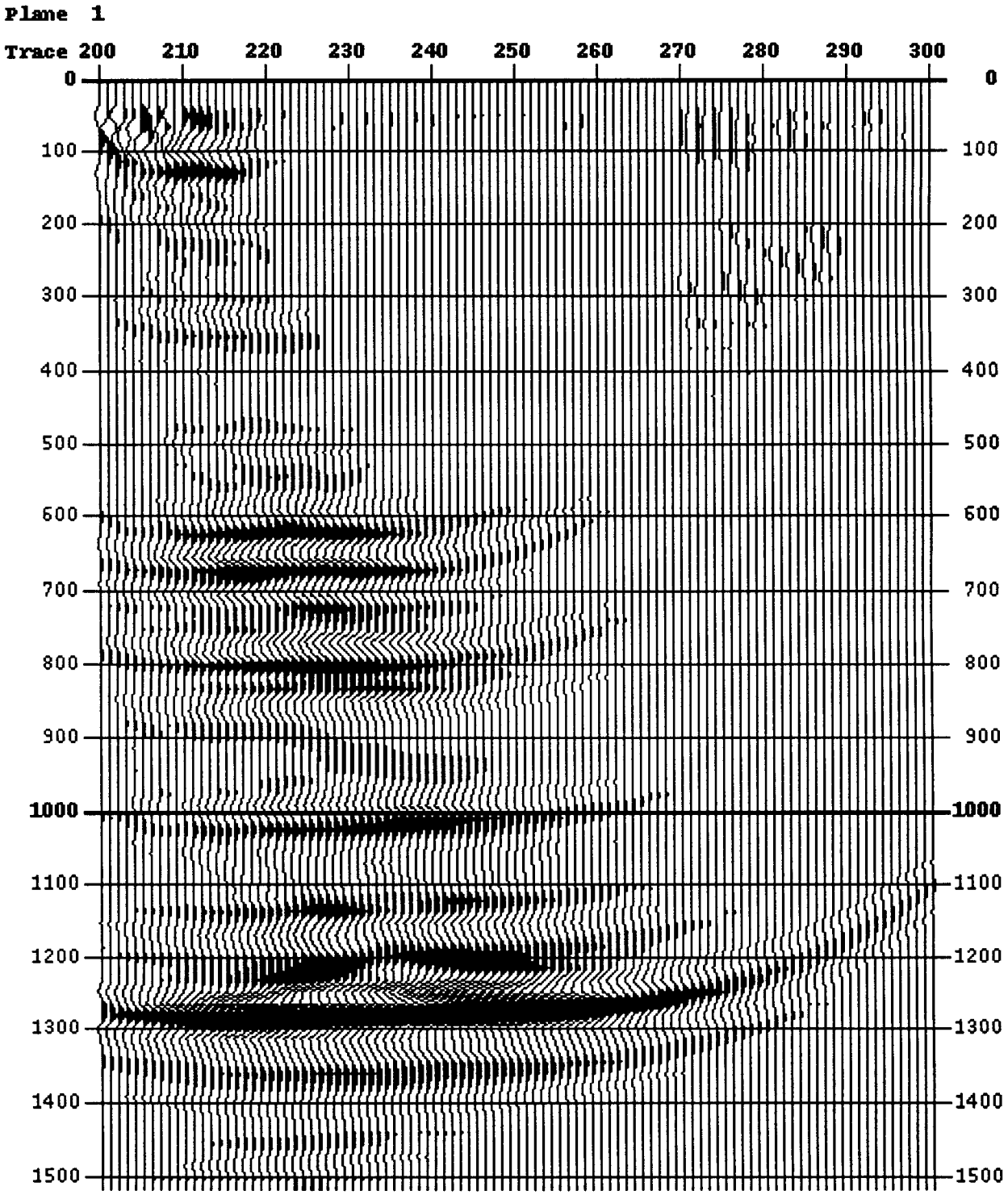

Plane 1
Trace
10
20
30
0
100
200
300
400
500
600
700
800
900
1000
1100
1200
1300
1400
1500
1600
1700
1800
1900


Plane 1
Trace
10
20
30
0
100
200
300
400
500
600
700
800
900
1000
1100
1200
1300
1400
1500
1600
1700
1800
1900

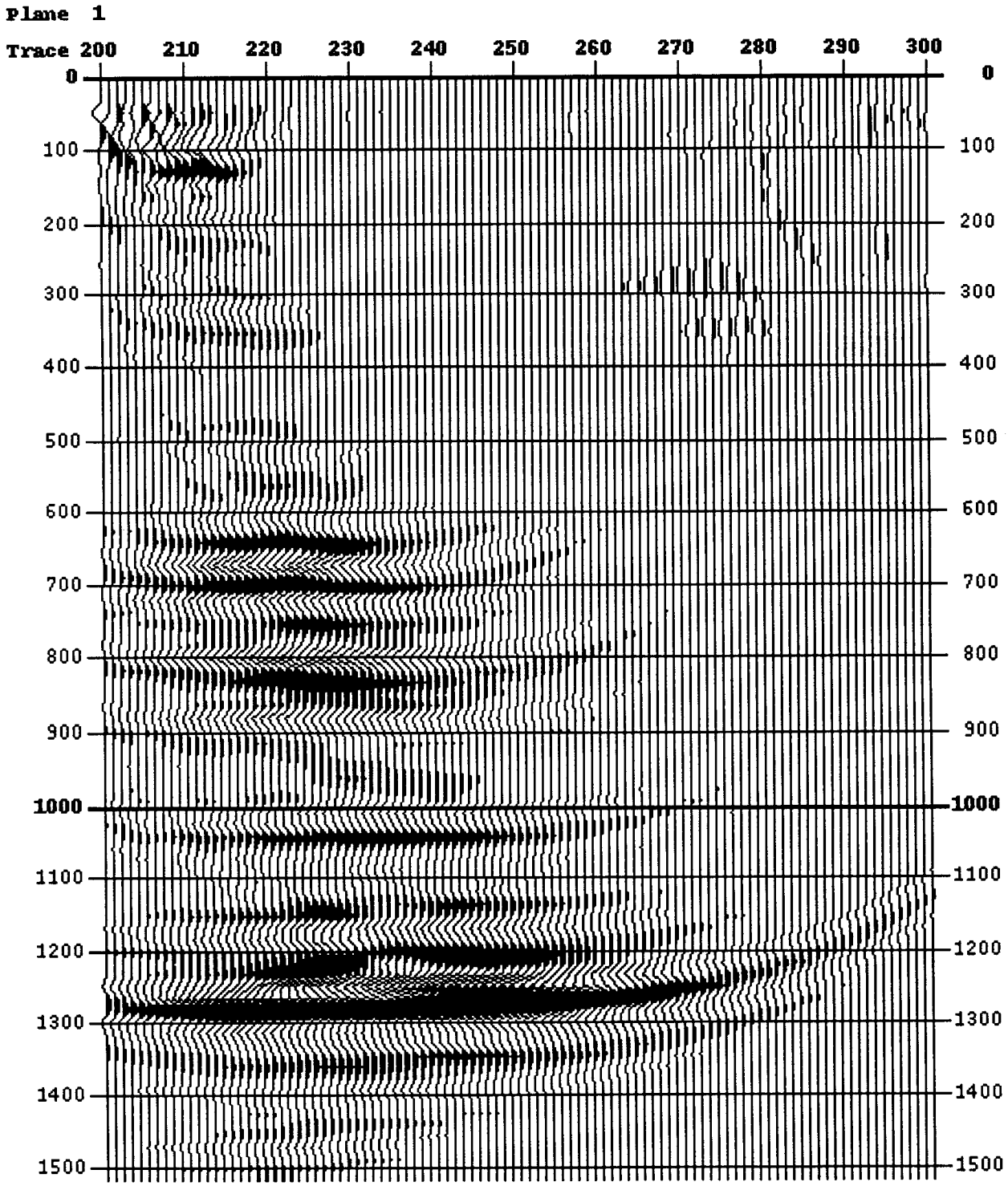

2
1.9
1.8
1.7
1.6
1.5
1.4
0
200
400
600
800
1000
1200
1400
1600
1800
2000


Plane 1
Trace 200 210 220 230 240 250 260 270 280 290 300
0 100 200 300 400 500 600 700 800 900 1000 1100 1200 1300 1400 1500

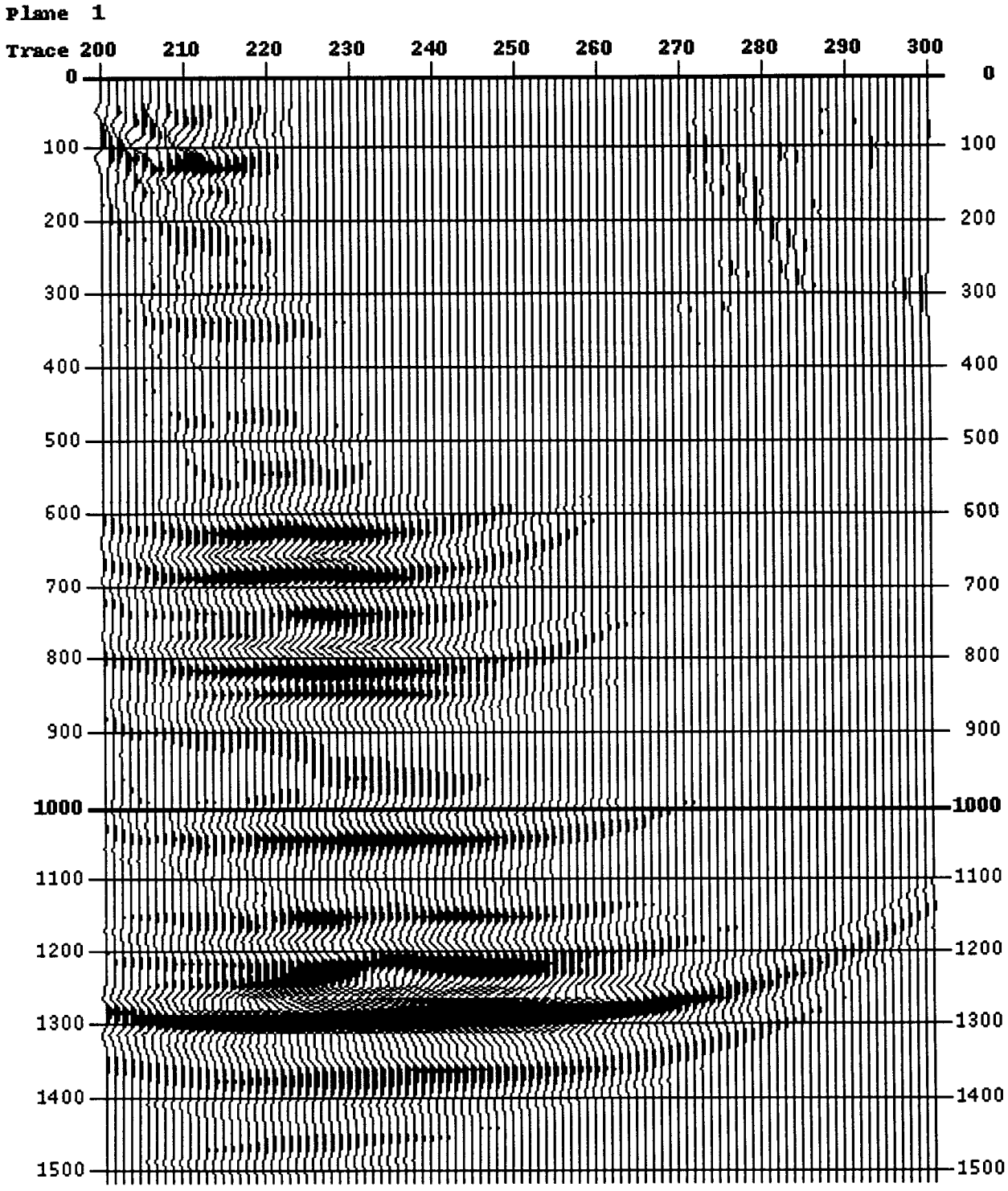

2
1.9
1.8
1.7
1.6
1.5
1.4
0
200
400
600
800
1000
1200
1400
1600
1800
2000


Plane 1
Trace 200 210 220 230 240 250 260 270 280 290 300
0 100 200 300 400 500 600 700 800 900 1000 1100 1200 1300 1400 1500

12a

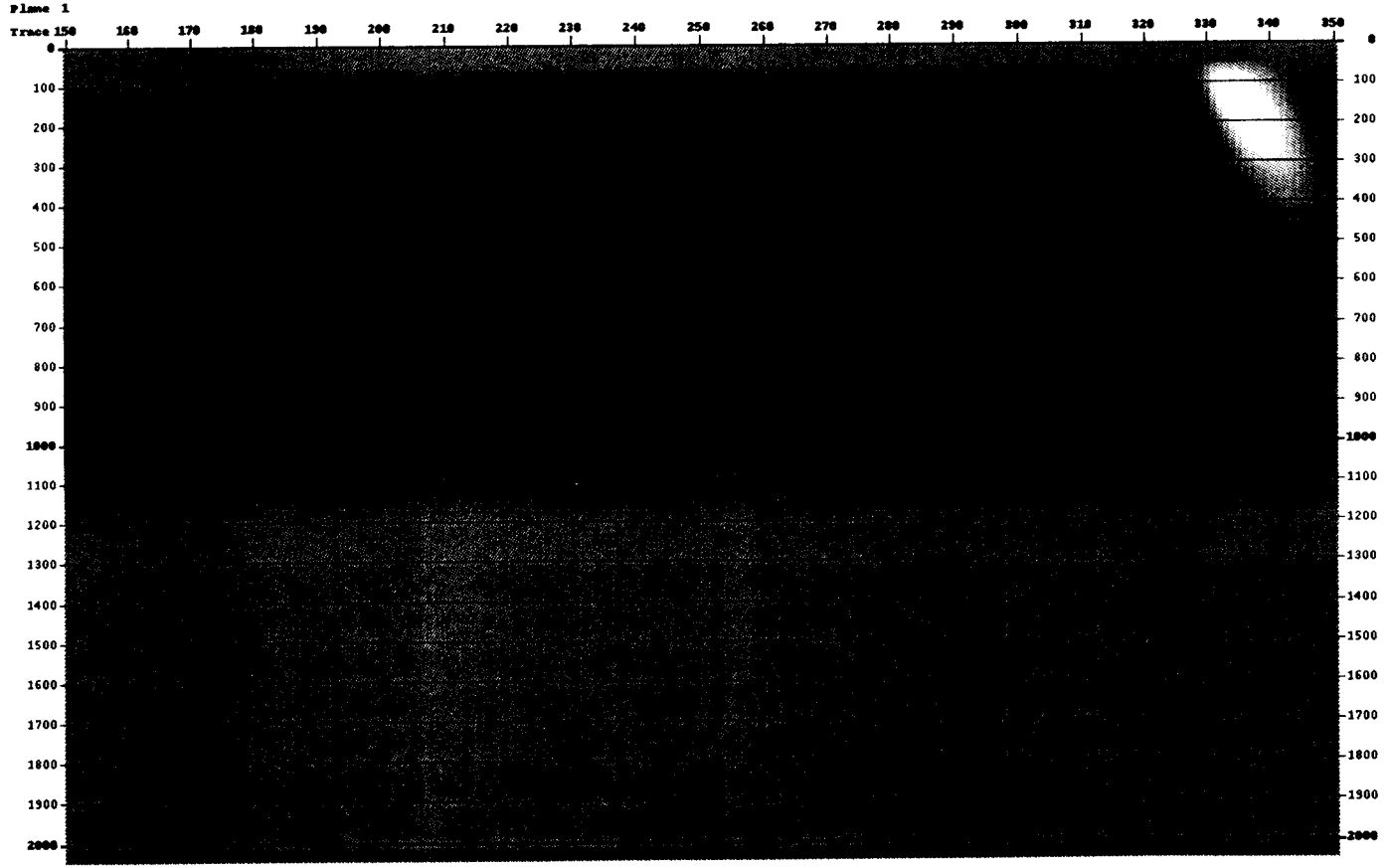

Plane 1
Trace 200 210 220 230 240 250 260 270 280 290 300
0
100
200
300
400
500
600
700
800
900
1000
1100
1200
1300
1400
1500

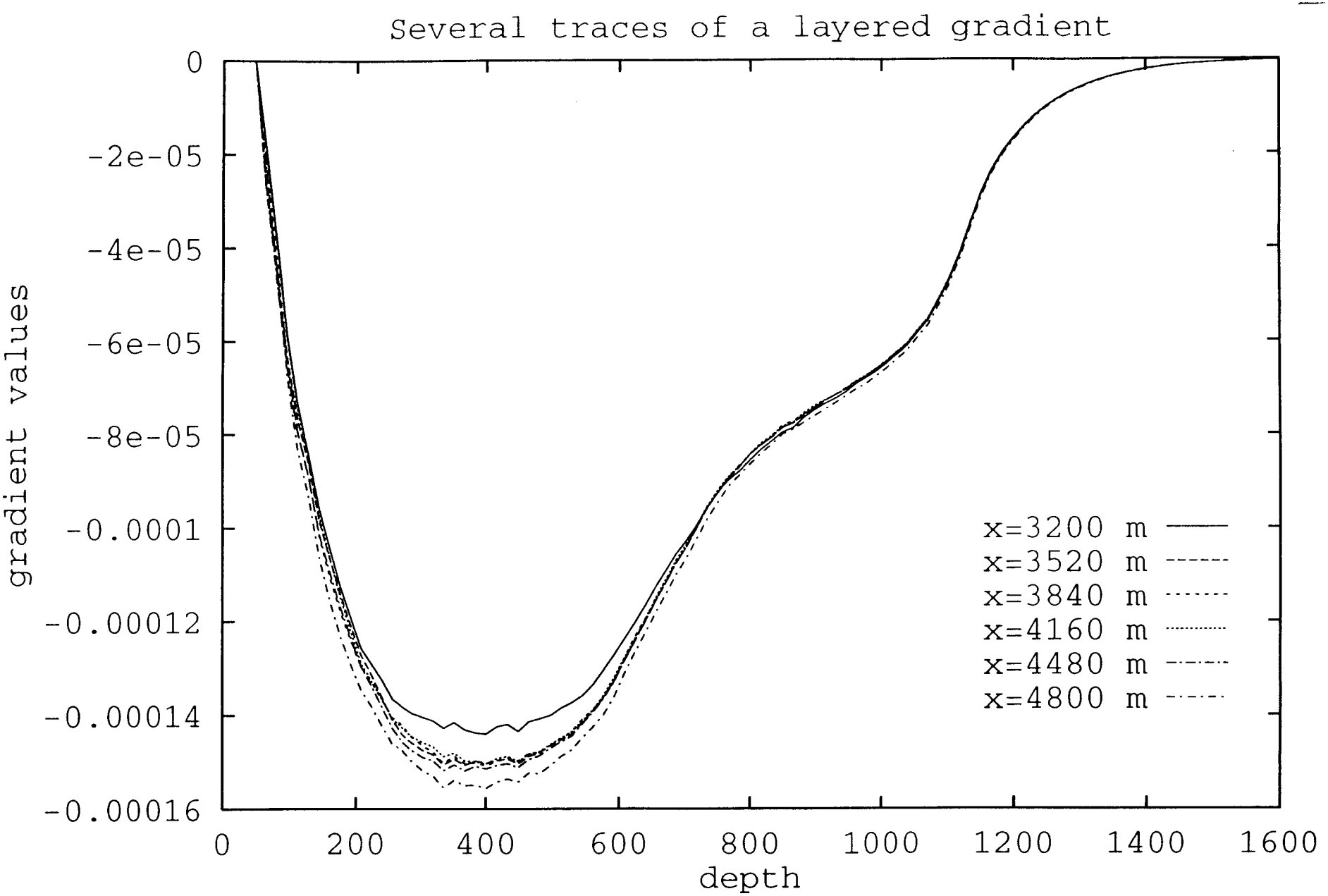
12b


Plane 1
Trace 200 210 220 230 240 250 260 270 280 290 300
0
100
200
300
400
500
600
700
800
900
1000
1100
1200
1300
1400
1500

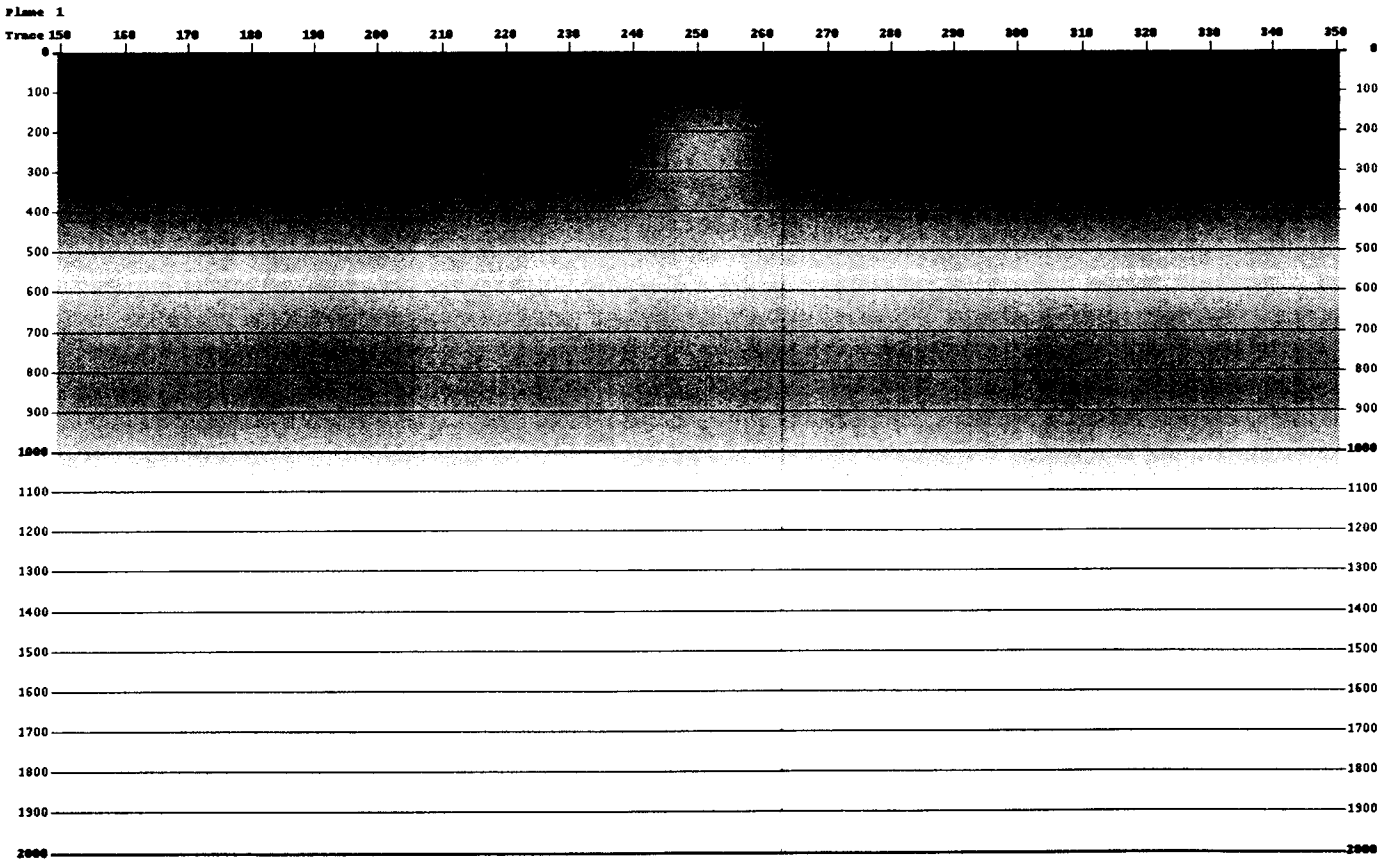
12c

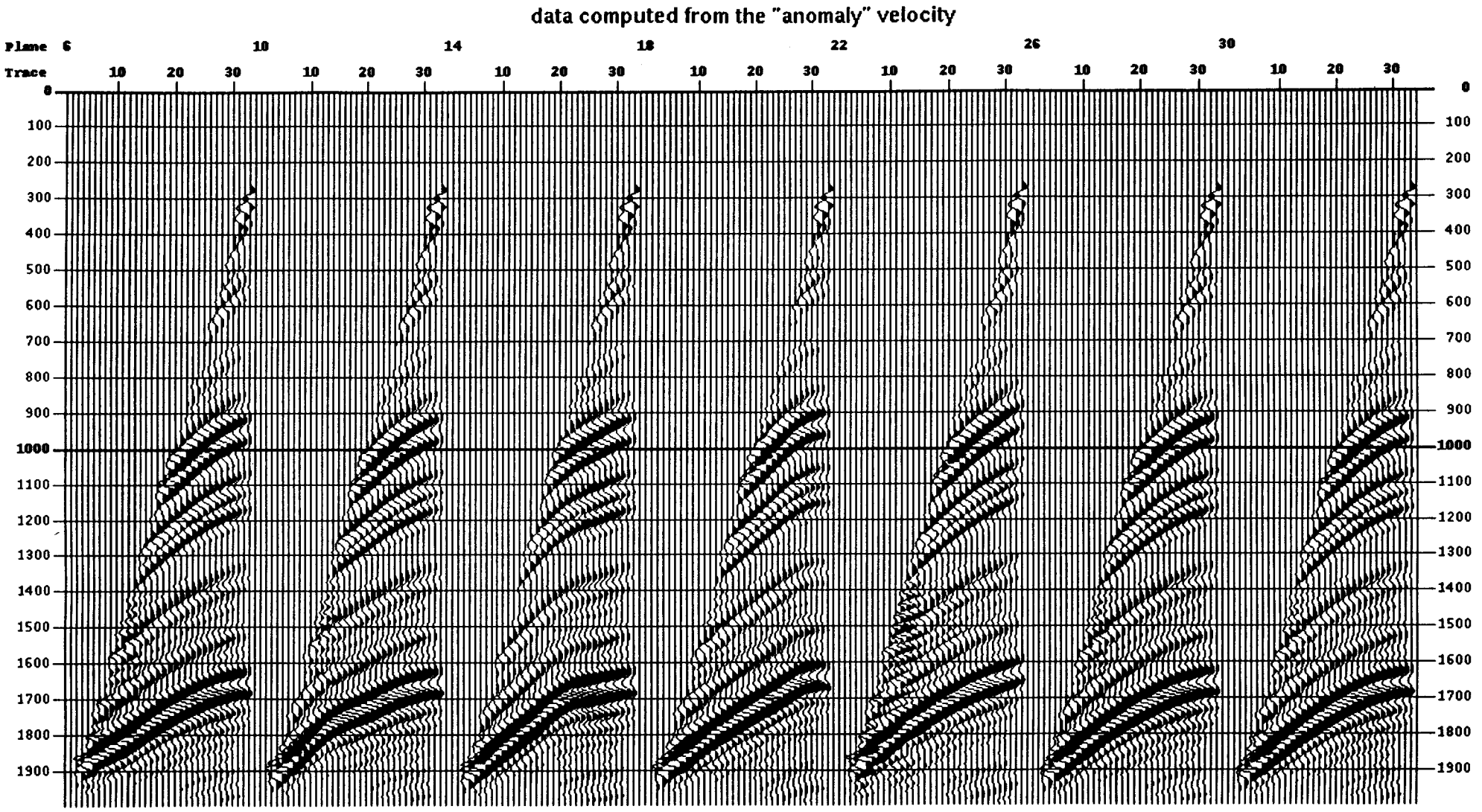
Gradient for layered data at constant velocity – Isotropic H1 100 m



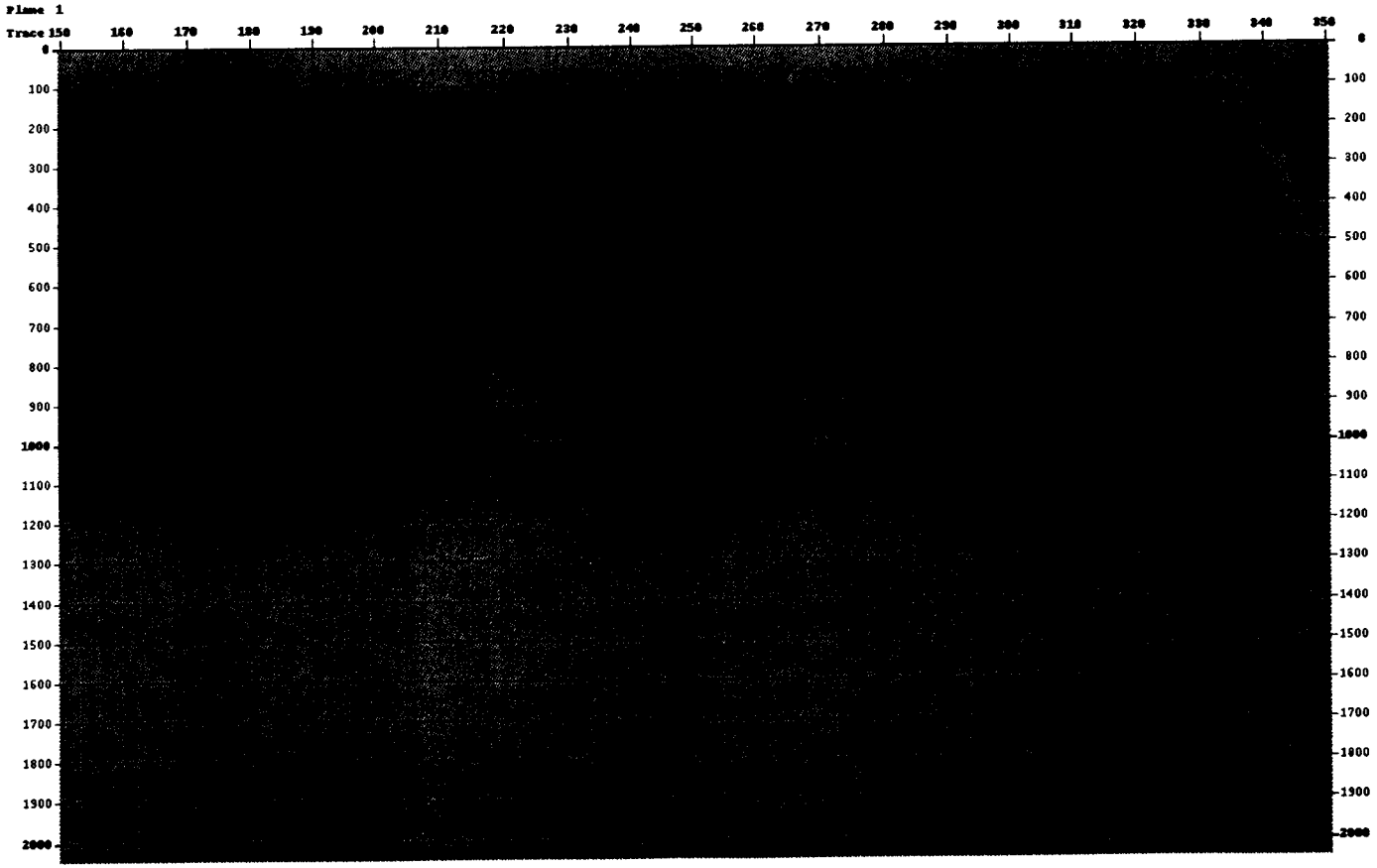

Several traces of a layered gradient
gradient values
depth
0
-2e-05
-4e-05
-6e-05
-8e-05
-0.0001
-0.00012
-0.00014
-0.00016
0
200
400
600
800
1000
1200
1400
1600
x=3200 m
x=3520 m
x=3840 m
x=4160 m
x=4480 m
x=4800 m

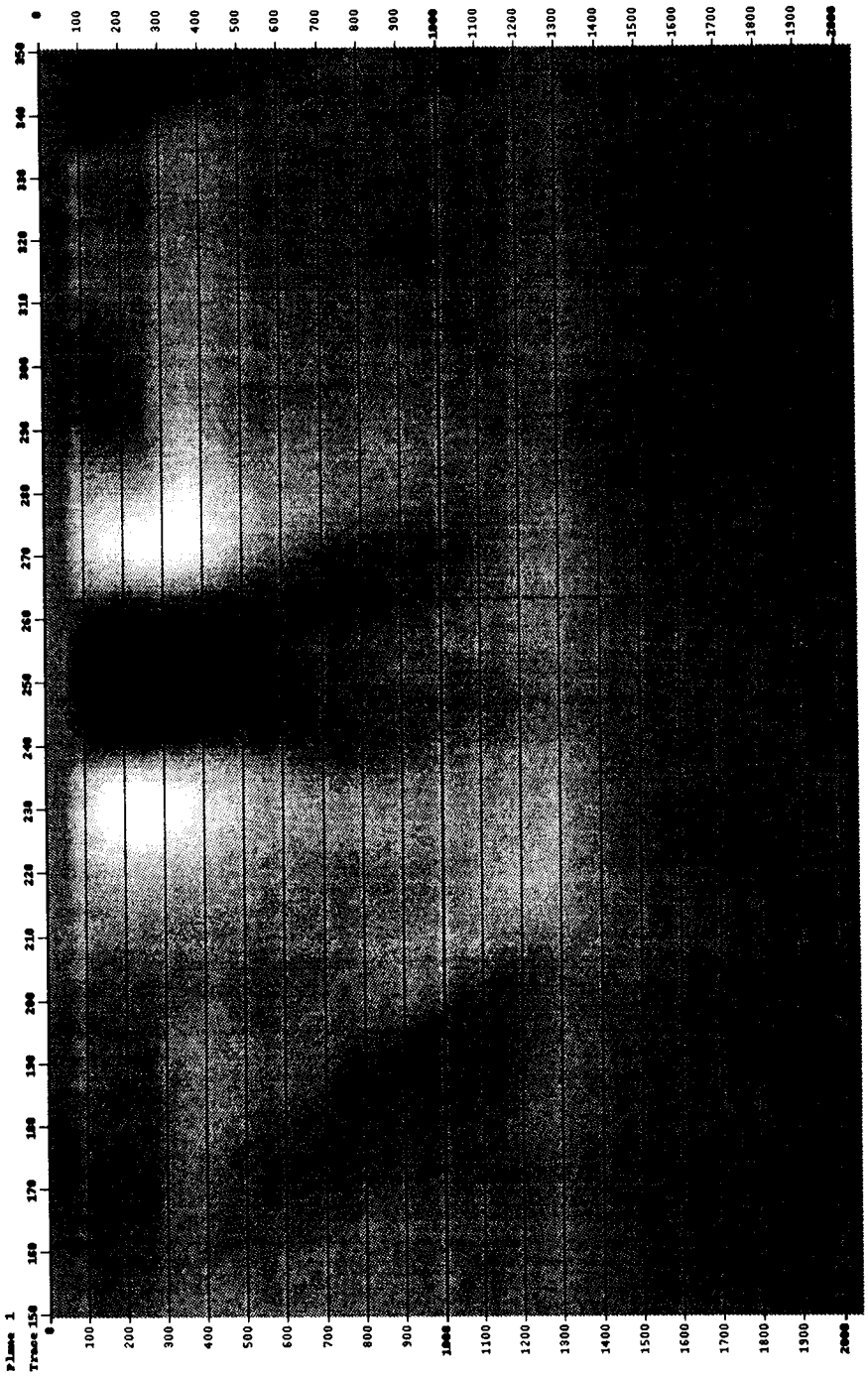
Velocity model with shallow high anomaly

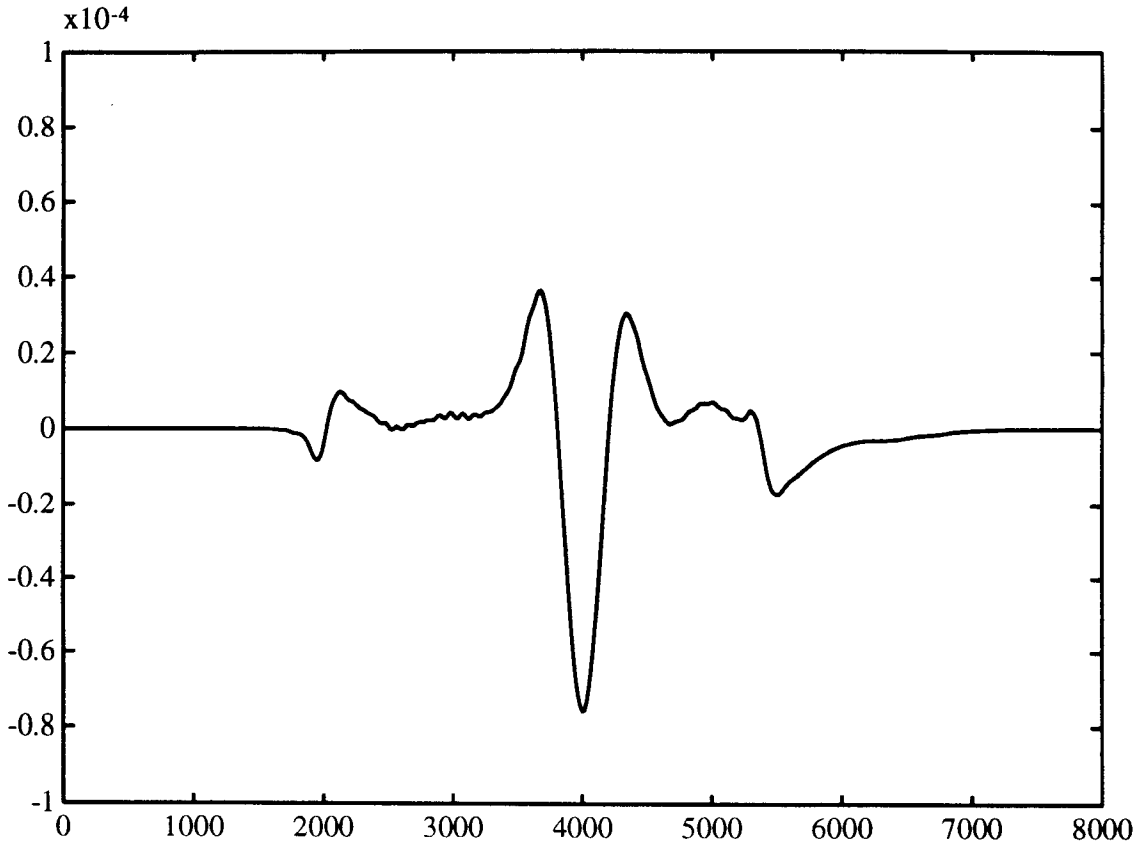



data computed from the "anomaly" velocity
Plane 6 10 14 18 22 26 30
Trace 10 20 30 10 20 30 10 20 30 10 20 30 10 20 30 10 20 30 10 20 30
0 100 200 300 400 500 600 700 800 900 1000 1100 1200 1300 1400 1500 1600 1700 1800 1900

Gradient at constant velocity – Isotropic H1 100 m




Gradient at layered velocity – Isotropic H1 100 m
Plane 1
Trace 150


x10-4
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
0
1000
2000
3000
4000
5000
6000
7000
8000